

LEARNING ACTIVITIES  
FROM HUMAN DEMONSTRATION VIDEOS

Jangwon Lee

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements for the degree

Doctor of Philosophy

in the School of Informatics, Computing, and Engineering,

Indiana University

November 2018

Accepted by the Graduate Faculty, Indiana University,  
in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

---

David J. Crandall, Ph.D.

---

Selma Šabanović, Ph.D.

---

Chen Yu, Ph.D.

---

Donald S. Williamson, Ph.D.

July 19, 2018

Copyright © 2018

Jangwon Lee

## ACKNOWLEDGMENTS

I always see the famous phrase “Stand on the shoulders of giants” by Isaac Newton when I open Google Scholar to search a paper for my research. One day, during the last period of my PhD journey, I opened Google Scholar just like always and saw the phrase again, and then I suddenly felt to say that “That’s totally true Newton, but you may not know how hard it is for a ordinary person like me to climb giant’s shoulder.”

Today, I would like to borrow the quote from Lou Gehrig to say, “I am the luckiest man on the face of the earth.” Thanks to a lot of people’s help, I was able to get this point.

First and foremost, I would like to thank my advisor David Crandall who climbed the giant’s shoulder in advance and lowered a rope for me. His guidance, patience, and wholehearted support helped me to stand up again when I wanted to give up the challenge. I would also like to express special thanks to my another advisor Selma Šabanović for her persistent help, support, encouragement and the opportunities she have provided me during my PhD journey with her amazing R-House lab. At least I think that I was the luckiest Ph.D student on earth without doubt about it because I had you two as my advisors.

In addition, I would like to thank Chen Yu and Donald Williamson for their time and efforts to serve on my research committee. I was able to get a lot of insight from your invaluable suggestions and feedback in discussions. I also want to say thanks to all the other faculty and staff of Indiana University, especially Informatics graduate office, for providing me the great opportunity to study at IU and helping me a lot with their expertise and time.

Thank you also to all of my friends and colleagues at IU for being good travel companions during my PhD journey. IU Computer Vision Lab and R-House Living lab members in particular. I would like down all of your names, but there is not enough space to in the margin of my dissertation to write it.

Finally, I thank my two little humans, Hana and Yuna, and my wife Eunyoung. I wish to know the word more than love to express my love whenever I feel like it. I would also like to thank to my parents and brother who are always on my side. You are the backbone of everything I do in my life.

Honestly, I still feel like I am the middle of climbing the giant's shoulder. Perhaps I am still around the waist of the giant, but now I believe that I can reach the giant's shoulder sometime in the future thanks to all of your help, my next door giants.

This thesis was based on work supported in part by the Air Force Office of Scientific Research under award FA9550-13-1-0225, by the National Science Foundation (CAREER IIS-1253549), and by the Indiana University Emerging Areas of Research (EAR) Program.

Jangwon Lee

LEARNING ACTIVITIES  
FROM HUMAN DEMONSTRATION VIDEOS

An important goal of intelligent systems is to carry out human-like actions, either alone or in collaboration with a person. Instead of requiring professional engineers to explicitly program these behaviors, systems could instead learn them automatically by observing videos of humans demonstrating the tasks. However, this is challenging due to the large number and variety of actions that people perform in the context of dynamic and uncontrolled real environments. For example, people may carry out the same action differently, or react differently to the same situation. Moreover, many human behaviors have complex movements that are difficult to observe accurately.

In this thesis we describe novel computer vision approaches to observe and learn activities from human demonstration videos. We specifically focus on using first-person and close up videos for learning new activities, rather than traditional third-person videos that have static and global fields of view. Since the specific objective of these studies is to build intelligent agents that can interact with people, these types of videos are beneficial for understanding human movements, because first-person and close up videos are generally goal-oriented and have similar viewpoints as those of intelligent agents. We present new Convolutional Neural Network (CNN) based approaches to learn the spatial/temporal structure of the demonstrated human actions, and use the learned structure and models to analyze human behaviors in new videos. We then demonstrate intelligent systems based on the proposed approaches in two contexts: (i) collaborative robot systems to assist users with daily tasks, and (ii) an educational scenario in which a system gives feedback on their movements. Finally, we experimentally evaluate our approach in enabling intelligent systems to

observe and learn from human demonstration videos.

---

David J. Crandall, Ph.D.

---

Selma Šabanović, Ph.D.

---

Chen Yu, Ph.D.

---

Donald S. Williamson, Ph.D.

# CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Thesis Overview . . . . .	1
1.2	Human Demonstation Videos . . . . .	3
1.3	Modeling Human Activities . . . . .	4
1.3.1	Human Behavior Understanding . . . . .	4
1.3.2	First-Person Videos . . . . .	7
1.3.3	Early Recognition . . . . .	9
1.4	Motivation and Problem Statement . . . . .	10
1.5	Summary and Thesis Outline . . . . .	11
<b>2</b>	<b>Learning New Tasks by Imitating Human Behaviors</b>	<b>14</b>
2.1	Robot Learning from Demonstration . . . . .	15
2.2	Learning Human-Robot Collaborative Tasks . . . . .	18
2.2.1	Intention Recognition . . . . .	19
2.2.2	Legibility of Robot Intention . . . . .	24
2.2.3	Interactive/Active Learning . . . . .	29
2.3	Summary of Related Work . . . . .	32
<b>3</b>	<b>Observing Human Behaviors in Demonstration Videos</b>	<b>36</b>
3.1	Analyzing Human Hands in Human-Human Interaction Videos . . . . .	36
3.1.1	Introduction . . . . .	36



3.1.2	Approach: Convolutional Future Regression . . . . .	38
3.1.3	Experimental Results . . . . .	43
3.2	Analyzing Objects in Videos of Daily Living . . . . .	47
3.2.1	Introduction . . . . .	47
3.2.2	Approach: Two-Stream Network with a Temporal Stream . . . . .	48
3.2.3	Experimental Results . . . . .	50
3.3	Analyzing Pressed Notes and Finger Movements of People Playing Piano .	54
3.3.1	Introduction . . . . .	54
3.3.2	Related Work . . . . .	56
3.3.3	Approach: Multi-Task Learning With Video-Audio Fusion . . . . .	58
3.3.4	Experimental Results . . . . .	63
<b>4</b>	<b>Building Interactive Systems for Natural Interaction with People</b>	<b>70</b>
4.1	From Perception to Intelligent Agents . . . . .	71
4.2	Forecasting Hand Gestures for Human-Drone Interaction . . . . .	72
4.2.1	Introduction . . . . .	72
4.2.2	Approach: Interactive Control based on Forecasting Hand Gestures	73
4.2.3	Experimental Results . . . . .	74
4.3	Human-Robot Collaboration with a Humanoid Robot . . . . .	76
4.3.1	Introduction . . . . .	76
4.3.2	Approach: Manipulation Network for Imitating Human Behaviors .	77
4.3.3	Experimental Results . . . . .	78
4.4	A Case Study for Building a Sociable Robot . . . . .	81
4.4.1	Introduction . . . . .	81
4.4.2	Approach: Card Sorting with Focus Group Interview . . . . .	82
4.4.3	Findings . . . . .	84

4.4.4	Discussion . . . . .	85
<b>5</b>	<b>Conclusion and Future Work</b>	<b>87</b>
5.1	Thesis Summary and Discussion . . . . .	87
5.2	Future Work . . . . .	89
	<b>Bibliography</b>	<b>91</b>
	<b>Curriculum Vitae</b>	

## LIST OF FIGURES

1.1	Action representation . . . . .	5
1.2	3D ConvNets for video analysis . . . . .	6
1.3	Two-stream network for video recognition . . . . .	7
2.1	Robot learning from demonstration . . . . .	15
2.2	Correspondence problem . . . . .	16
2.3	Affordance heat maps . . . . .	22
2.4	Legibility and Predictability of Robot Motion . . . . .	27
2.5	Active Learning . . . . .	30
3.1	Perception component for analyzing human hands . . . . .	38
3.2	Data flow of our perception component for analyzing the movements of hands	40
3.3	Two examples of our visual prediction . . . . .	46
3.4	Two-stream network with a temporal stream . . . . .	49
3.5	Object location forecasts on the ADL dataset . . . . .	53
3.6	Example object location forecasts in two different time setting . . . . .	53
3.7	Outline of our two-stream architecture for observing people playing a piano	58
3.8	Pipeline to create the piano dataset . . . . .	64
3.9	Our piano room with an experimental setup and an example of Hanon Exercises	65
3.10	Used fingers identification to press piano notes . . . . .	68
4.1	Data flow of our early gesture recognition systems . . . . .	73

4.2	Hand gestures for ‘come’ and ‘go’ to interact with a drone . . . . .	74
4.3	Confusion matrix of our gesture forecasting . . . . .	75
4.4	Robot manipulation component of our approach . . . . .	77
4.5	Qualitative results of our real-time robot experiments . . . . .	81
4.6	A drone delivery service . . . . .	83
4.7	A card sorting . . . . .	84

## LIST OF TABLES

3.1	Evaluation of future hand locations prediction . . . . .	44
3.2	Mean pixel distance between ground truth and predictions . . . . .	45
3.3	Mean pixel distance between ground truth and predicted position of right hand	45
3.4	Evaluation of future object presence forecasting . . . . .	51
3.5	Evaluation of future object location forecasting . . . . .	52
3.6	Evaluation of future hand locations prediction with two-stream network . .	54
3.7	Pressed Notes Detection Accuracy on One Hand Hanon dataset . . . . .	66
3.8	Pressed Notes Detection Accuracy on Two Hands Hanon dataset . . . . .	67
4.1	Real-time robot experiment results with human subjects . . . . .	80

## CHAPTER 1

### INTRODUCTION

#### 1.1 THESIS OVERVIEW

We often learn new skills for various reasons, from moving into a different field of work, to staying up-to-date in a rapidly changing world, to having new and different experiences, or to learning something new just for fun. Observing and imitating a teacher’s demonstration is often the best way to learn new skills. However, sometimes we are not able to have an instructor with us. In such cases, watching a teacher’s demonstration videos can be an effective alternative, because it does not constrain geographical location or time. Moreover, it has the additional benefit of being more student-centered, so that students can choose the most relevant videos and learn at their own pace.

This thesis is based on the hypothesis that a learning strategy incorporating video demonstrations can also be beneficial for intelligent agents (i.e., robots). If an intelligent agent could autonomously learn a new skill by watching demonstrations itself, it would not require human experts to explicitly program the skills nor would it require human teachers during the learning process. This avoids human labor to repeat the same action multiple times (potentially hundreds or thousands of times), for example. Furthermore, this learning style enables end-users without any professional engineering background to teach intelligent agents.

In this thesis, we propose to use human demonstration videos as a teaching resource for

training intelligent systems to interact with people and their environment. To be specific, we focus on using first-person or close-up videos for teaching intelligent agents new behaviors, rather than using traditional third-person videos. This helps the intelligent agent learn the teacher’s actual behaviors, because it places the agent in the teacher’s perspective; this type of video implicitly captures the teacher’s own field of view and encodes attention and intention of their behaviors. Therefore, the intelligent agent can learn where to look to imitate the teacher’s behavior and how to react according to its visual input thanks to the similarity of its field of view to that of the first-person videos. Nevertheless, this approach has some challenges such as low-quality videos due to camera motion, and the difficulty of capturing the overall intention of the teacher’s behaviors: since the viewpoint prevents us from observing the teacher’s whole body and their surrounding environment. However, this type of video tends to be more action-focused, compared to third-person videos which tend to have a global view of the scene, we thus can learn more fine-grained actions.

This dissertation presents new Convolutional Neural Network (CNN) based approaches for learning new skills from first-person (or action focused) human demonstration videos. These are data-driven approaches that rely on training datasets, so we first collect human demonstration videos according to various application scenarios. We consider two specific applications of intelligent systems: (i) interactive robots for helping users with daily tasks, and (ii) tutoring systems for giving feedback on a learner’s performance. We then introduce new models to capture the spatio-temporal structure of the demonstrated human actions in the videos. Next, we present novel approaches for transferring the learned skills from the demonstrations to robot actions, to help intelligent systems respond appropriately to their sensory input and application scenarios. After that, we evaluate the proposed approaches in terms of perception accuracy on test sets of collected demonstration videos. Finally, we evaluate how well intelligent agents successfully learn the desired behaviors through user

studies.

## 1.2 HUMAN DEMONSTATION VIDEOS

In this thesis, we use human demonstration videos to learn various activities for building intelligent and interactive systems. Our learning approach is based on the research paradigm of robot learning from demonstration (LfD), which attempts to enable robots to autonomously learn new tasks from observations of human demonstration. This concept has recently attracted interest in robotics (and particularly for service robots) because it allows end-users to teach robots by simply showing how to perform tasks that fit their needs in their own environment, on demand [5, 9]. Another advantage of LfD is that it allows intelligent agents to learn new tasks in novel situations faster than traditional engineering approaches, which require manually programming each task.

However, there are drawbacks associated with the use of LfD in the field of robotics. First, little attention has been paid to considering how to learn new tasks without real physical robots. Most previous studies require hardware platforms to learn control policies for the tasks from robot demonstrations, which makes it impractical for use by everyday concerns well who do not have this equipment. Moreover, previous work also often requires that end users have substantial expertise with the control systems of the robots, that which also contradicts the main principle of LfD. Third, learned knowledge is mostly hardware dependent, and therefore difficult to transfer to different types of robots. Finally, previous research tends to overlook the amount of time required for end-users to teach new tasks to their robots. They usually need to show the same demonstration multiple times to generalize each task.

This dissertation proposes a new methodology for learning activities from human demonstration videos without a physical platform like a real robot. This learning strategy is hard-



ware independent, and thus the learned activities can be transferred to different types of intelligent systems with proper scene and task representations. In addition, end-users of the system do not need to be physically co-present with the robots for teaching new tasks.

There are practical challenges and constraints for training intelligent agents solely based on human demonstration videos. For instance, new learning methods are required to adopt the learned knowledge from demonstration videos for providing physical services. However, this research can draw from, and generate fresh insight into, the related fields of artificial intelligence to use demonstration videos as training resources for learning human activities.

### **1.3 MODELING HUMAN ACTIVITIES**

#### **1.3.1 HUMAN BEHAVIOR UNDERSTANDING**

Understanding human behaviors in videos is essential for building a wide range of intelligent systems such as autonomous vehicles, intelligent surveillance systems, automatic analysis software for sports, etc. A considerable amount of literature has been published on human behavior recognition and analysis in the computer vision community [11,51]. This is a broad topic and must not only consider action recognition like detecting gestures and interactions, but also scene context. Nevertheless, as a starting point, in this thesis we focus on human action recognition to understand and learn the activities from videos. For this reason, in this section, we briefly introduce some major human action recognition approaches among the topics of human behavior understanding.

#### **Human Action Recognition**

A key aspect of recognizing human actions in videos is how we model the spatio-temporal structure of human movement. Thus a variety of recognition methods have been suggested to derive the spatio-temporal structure of an event, from low-level appearance based ap-

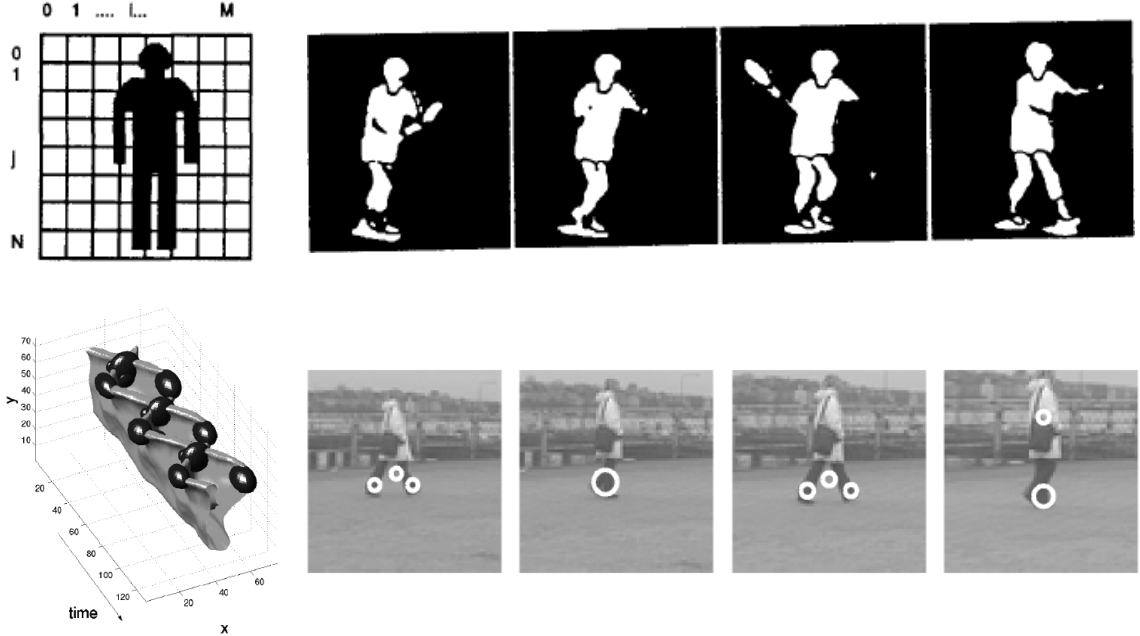


Figure 1.1: Extracting the spatio-temporal structure of events in videos. Top row shows the mesh feature and a sequence of extracted features which are used in [119]. Bottom row shows STIPs features and detected features in a sequence of image in [59].

proaches (e.g., Motion History Images (MHI) [10] and Spatio-Temporal Interest Points (STIPs) [59]) to high-level representation based models (e.g., using long-term trajectories [95], Bag of Correlated Poses (BoCPs) [117], and hierarchical approaches like Context Free Grammars (CFGs) on top of low-level features [48]).

For example, Yamato *et al.* published pioneering work in 1992 which used the ratio of the number of black pixels to the background white pixels in a mesh as a scene representation [119]. After extracting human pixels from the background image, each frame of video is then interpreted as a symbol based on the scene representation. Finally, they applied Hidden Markov Models (HMMs) on top of a sequence of those symbols to recognize videos of people playing tennis. In 2003, Laptev *et al.* reported a new feature to extract spatio-temporal interest points that encode temporal information jointly with spatial features [59]. It is designed to find a high variation of image intensity in  $x$ ,  $y$ , and *time* dimensions in a video, which can be interpreted as a video version of local features like the Scale-Invariant Feature

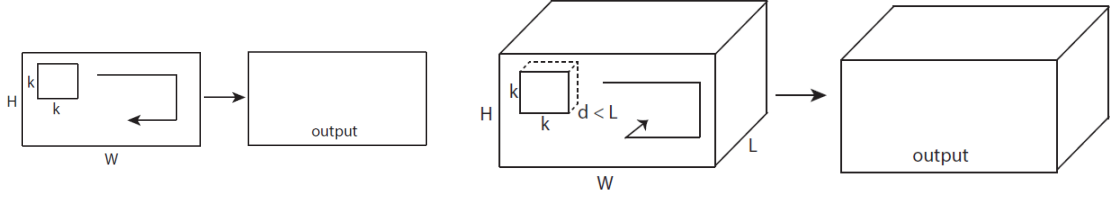


Figure 1.2: The difference between 2D convolution (left) and 3D convolution operations (right) in [112]. The output of 3D convolution operation is another 3D volume that contains temporal information about input videos.

Transform (SIFT) [73] which attempts to find interesting key-points in an image. They applied this feature for pose estimation of walking people in an outdoor scene.

More recently, literature has emerged that applies deep learning based approaches to represent the spatio-temporal structure of human behaviors [49, 112, 116]. One study by Du Tran *et al.* introduced a new deep learning based feature for video analysis [112]. Instead of applying a traditional 2D based convolutional network on videos, they suggested a 3-dimensional convolutional network to capture temporal information along with spatial signals. Fig. 1.2 illustrates the difference between 2D convolution and 3D convolution operations; the 3D convolution network computes convolution operations in time as well as space, so the output of network has a 3D shape. Another key approach to deep learning based action recognition methods for videos uses two different networks to extract spatial and temporal information separately, and then the outputs of two separate networks are combined into the final feature vector which contains both spatial and temporal signals. In 2014, Karen and Andrew showed this two-stream architecture for action recognition yielded large performance improvement, and had a major impact on activity recognition in the field [103].

In this thesis, we introduce novel convolutional neural networks to capture the spatio-temporal structure of human actions, which we would like to model to allow intelligent

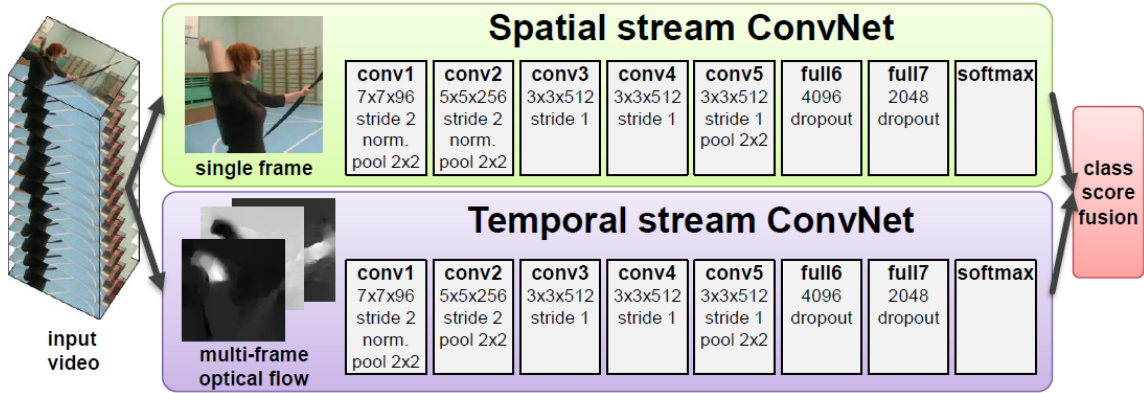


Figure 1.3: Two-stream architecture of Karen and Andrew [103]. While spatial stream ConvNet captures spatial information individual from still frames of input video, temporal stream ConvNet extracts motion information between frames.

systems to understand and imitate human behaviors. Our architectures are influenced by existing work described in this section, particularly deep learning based approaches; however, our approaches are designed to go beyond recognizing human actions. Since the central thesis of this dissertation is building intelligent agents that can interact with the environment and people, we also consider minimizing the transfer cost to hand over learned actions to intelligent systems. For this reason, we use first-person (or action focused) videos as our main training resource and design our network for handling this particular type of video. With this, we can reduce the knowledge transfer cost because this type of video has a similar viewpoint as the intelligent systems. Moreover, we also investigate how to generate feedback from perception of human actions based on our proposed approaches. Since systematic studies of how human activity recognition contributes to an intelligent agent (i.e., a robot) or how those behaviors can be learned and reproduced are still lacking, we believe this thesis offers important insights into the field of computer vision and robotics.

### 1.3.2 FIRST-PERSON VIDEOS

First-person or egocentric vision is a sub-field of computer vision that analyzes images and videos that are taken by wearable cameras such as GoPro and Google Glass. Since they

show how the camera wearer perceives and interacts with the world from his/her point of view, these cameras have unique properties because they implicitly contain the camera wearer’s intentions and experiences.

First-person videos thus have attracted interest from computer vision researchers because of their increasing popularity and unique properties, and a considerable literature has grown around the egocentric vision in recent years. This includes recognition of ego-actions (the camera wearer’s actions) like jumping and turning in sports videos [52], understanding daily activities such as making coffee [32], and discovering important events in long and unstructured egocentric videos [74].

Research into egocentric vision is also relevant to robotics researchers because robots perceive the world from a first-person point of view as well. Consequently, robots can take advantage of computer vision approaches developed in the field of egocentric vision. For example, a recent study by Gupta *et al.* (2017) showed how to map first-person robot camera inputs to corresponding navigational actions for robot visual navigation [41]. The core idea of this paper is to learn how to move a robot (planning) jointly with a spatial memory that corresponds to an egocentric input image (mapping) based on an autoencoder neural network. So far, however, most of the proposed approaches in the field are limited to specific computer vision problems like activity classification, object detection, and object/activity segmentation [32, 74, 94]

In this thesis we use wearable cameras to collect human activity videos in order to take advantage of the information in first-person perspectives, such as the relationship between manipulated objects and the camera wearer’s hands, and information about which objects are the most relevant to the current activity of the camera wearer. Since this type of camera captures what the camera wearer sees during their activities, we can also track their attention and learn what he/she looks at before and after taking actions. Compared with

learning from the traditional third-person videos, we might miss some useful information such as facial expressions of the camera wearer and overall intention of their behaviors if we learn solely from egocentric videos. However, this type of video is beneficial to build intelligent systems that react according to their visual input because of the viewpoint similarity we described before, thus reducing the cost of knowledge transfer.

### 1.3.3 EARLY RECOGNITION

Another important line of related work to this thesis is early recognition, which aims to anticipate future events based on current observations and prior knowledge. Humans often have an ability to accurately estimate what will happen next given what we have observed so far. For instance, we can predict that two people will shake hands if they are approaching each other with outstretched arms. Likewise, the objective of this line of research is to give a similar ability to intelligent systems, since it will likely help create more natural human-machine interaction. There has been an increasing interest in early recognition research and several studies have investigated forecasting a future event as early as possible [53, 99, 114, 115].

For example, researchers have attempted to find the key elements of an ongoing event that affects the future state of the world [99, 115], and to integrate visual features with prior knowledge of the event for solving a problem [53]. More recently, some studies have investigated learning visual representations of future frames [114], or directly predicting future image frames based on deep learning algorithms [72].

However, previous published studies are limited to visual prediction systems that assume a static camera, not a first-person camera in which the camera wearer’s actions physically change the environment and/or objects of the scene. There exists recent work that considers the dynamics of first-person video [39, 98] and generates robot control actions based on

visual prediction [36], which shows the potential in applying visual prediction for building intelligent systems. Despite this, very few studies have investigated using early recognition approaches for robotics applications.

This dissertation proposes an early recognition approach with novel convolutional regression networks not only to predict future events but also to forecast the future locations and status changes of interesting objects. A major advantage of object location forecasting is that it enables intelligent agents to take actions without any additional processing like motion planning, since this forecasting ability provides the target position of all related objects and manipulators (hands) that the agent should move. Furthermore, forecasting would allow us to build intelligent tutoring systems which give hints or cautions to a learner in advance. The proposed approaches are particularly useful in terms of training the deep network since they do not require any additional annotations.

## 1.4 MOTIVATION AND PROBLEM STATEMENT

This dissertation aims to build intelligent systems that can automatically learn desired skills from observations of human demonstration videos. This is not just about making intelligent systems recognize human behaviors, but also about making them understand circumstances and reproduce them, which requires learning far more. More concretely, given human demonstration videos, our goal is (i) to analyze human behaviors (from coarse-grained to fine-grained) and all interactive objects in view of the videos, (ii) to learn human behaviors and transfer knowledge to the intelligent agent so that it can understand and imitate the demonstrated behaviors, and then (iii) to provide suitable responses to a human user's behaviors according to the application scenario.

Understanding human behaviors at a certain level is essential for building intelligent systems that interact with humans, and a considerable amount of literature has been pub-

lished on human activity recognition [60,112]. However, this problem is challenging because of uncertainty in human behaviors. Every person acts differently, even for the same task. For example, when making a sandwich, one person might first pick up a knife to spread butter on the bread, while another might first prepare the plate. In addition to uncertainty in human behaviors, the behaviors can look different across each attempt for the same task, due to environmental factors like appearance of the person, background changes, object location changes, lighting differences, viewpoint changes, and so on. Moreover, here we are concerned about understanding human behaviors at an executable level that intelligent agents can use to mimic the learned behaviors. This problem becomes much more challenging since it requires a bridge to connect sensory input with control commands of the intelligent agent.

In this thesis, we address the above challenges using new “deep learning” approaches with Convolutional Neural Networks. Recently, deep learning based approaches have delivered striking performance increases on a range of machine learning problems [67,97], but few studies have investigated deep learning based techniques for building intelligent systems that can interact with people. Therefore, this study makes a major contribution to research on activity learning for intelligent systems by showing how to enable intelligent agents to learn desired activities from human demonstration videos.

## 1.5 SUMMARY AND THESIS OUTLINE

The remainder of the thesis is composed of four themed chapters. The following chapter provides background and related work on learning new skills by imitating human behavior, focusing on human-robot collaborative tasks with computer vision approaches. In Chapter 3 we address the perception part for intelligent systems to understand and learn human behaviors from demonstration videos. We present novel CNN based computer vision ap-



proaches to detect human hands, gestures, and the objects with which people interact, and then demonstrate that our approaches can achieve promising results in terms of detection accuracy. Chapter 4 is concerned with the methodology used for transferring learned knowledge from the human demonstration videos to intelligent systems. We show how to connect the perception component of the intelligent systems with a feedback component. We then present our user study design and findings based on experimental results involving human participants. To conclude, we summarize the proposed dissertation and discuss open research questions for future work in Chapter 5.

### **Our published papers and their corresponding chapters:**

- Chapter 3.1
  - Jangwon Lee and Michael S. Ryoo, Learning Robot Activities from First-Person Human Videos Using Convolutional Future Regression, in the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017 [62]
- Chapter 3.2
  - Chenyou Fan, Jangwon Lee, and Michael S. Ryoo, Forecasting Hand and Object Locations in Future Frames, submitted to European Conference on Computer Vision (ECCV) Workshop on Anticipating Human Behavior, 2018 [30]
- Chapter 4.2
  - Jangwon Lee, Haodan Tan, Selma Šabanović, and David Crandall, Forecasting Hand Gestures for Human-Drone Interaction. in the ACM/IEEE International Conference on Human-Robot Interaction (HRI), Late-Breaking Reports, 2018 [63]
- Chapter 4.3
  - Jangwon Lee and Michael S. Ryoo, Learning Robot Activities from First-Person Human Videos Using Convolutional Future Regression, in the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017 [62]

- Chapter 4.4
  - Haodan Tan, Jangwon Lee, and Gege Gao, Human-Drone Interaction: Drone Delivery & Services for Social Events, in the Conference on Designing Interactive Systems (DIS), Work-in-Progress, 2018 [109]

**Other published papers not related to this thesis:**

- Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David Crandall, and Michael S. Ryoo, Identifying First-person Camera Wearers in Third-person Videos, in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 [31]
- Jangwon Lee, Jingya Wang, David Crandall, Selma Šabanović, and Geoffrey Fox, Real-Time, Cloud-Based Object Detection for Unmanned Aerial Vehicles, in the IEEE International Conference on Robotic Computing (IRC), 2017 [64]

## CHAPTER 2

### LEARNING NEW TASKS BY IMITATING HUMAN BEHAVIORS

The research paradigm of robot Learning from Demonstration (LfD) or robot Programming by Demonstration (PbD) is adopted in this dissertation to enable intelligent systems to autonomously learn new tasks by watching human demonstration videos. The benefit of this approach is that it enables non-robotics experts to teach robots without any professional background like mechanical engineering or computer programming. This research paradigm thus can play an important role in addressing the issue of scaling up robot learning. Existing research recognizes this attractive feature of LfD, so there is a growing body of literature that employs the theme of LfD [5, 9].

LfD is a broad topic ranging from various machine learning techniques like supervised learning, reinforcement learning, and feature selection, to human factors as well. Researchers with different backgrounds thus employ the concept from different points of view. However, there are common theoretical issues in the field such as the *Correspondence Problem* that arises due to a mismatch between the teacher’s body configuration and the student’s configuration [84], and the interface challenge of designing user-friendly interfaces for demonstration (i.e., motion-capture systems [123]) in order to enable non-robotics experts to teach robots new knowledge without any difficulty. In this chapter, we first present a brief overview of LfD approaches, and then review recent research in this area while focusing on studies for human-robot collaborative tasks.

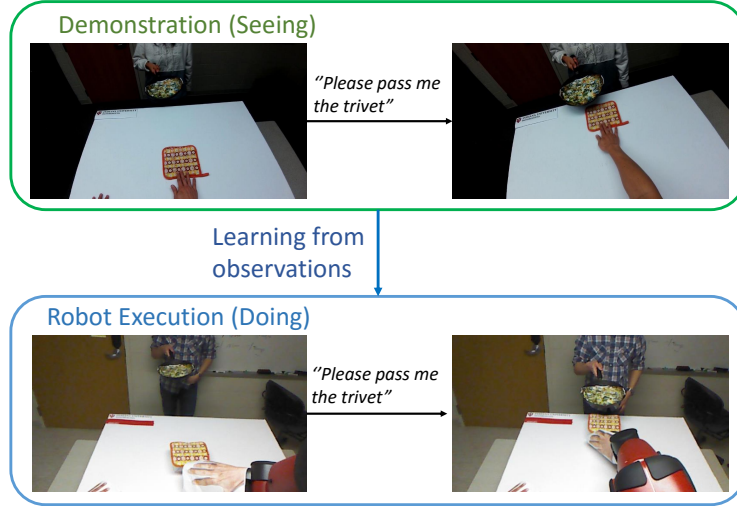


Figure 2.1: Robot learning from demonstration (LfD) enables robots to automatically learn a new task from observations. Ideally, end-users can program robots for a new task without any background of robotics technology and programming.

## 2.1 ROBOT LEARNING FROM DEMONSTRATION

LfD is a promising approach may help move robot prototypes from research laboratories to the real world, since it typically does not require any expert knowledge of robotics technology to teach robots new tasks. LfD thus allows end-users to teach robots what robots should do based on user’s own requirements at their place and opens up new possibilities of robot learning. For instance, a robot can learn new tasks that are infeasible to be learned using pre-programming like preparing a cup of coffee according to personal preferences. Recently, a considerable literature has grown around the theme of LfD because of this attractive feature [5, 9, 20]. However, LfD is also challenging because there are fundamental theoretical issues in this research paradigm. One challenge is called the *Correspondence Problem* that arises due to the different body configurations between humans and robots. The *Generalization Issue* also happens for learning a desired skill across a set of several different demonstrations. We now discuss these two problems in detail.

The Correspondence Problem is related to the question of how to reproduce a learned skill with a real robot, and thus depends on how a training dataset is recorded. Histor-



Figure 2.2: Illustration of the Correspondence problem: how to reproduce a learned skill with a real robot? This problem arises due to a mismatch between the teacher’s (human’s) body configuration and the student’s (i.e, robots) configuration. This figure is borrowed from [55].

ically, LfD research started with a memory-based (playback) method that simply records the position of the robot end-effector using an actual robot and reproduces the recorded positions with the same robotic platform [9]. Although we can avoid the correspondence problem if we gather the training examples using the same robot that is used for execution, we often would like to use human generated data as training examples. In such cases, we need to consider how the human generated demonstrations can be encoded and how they are interpreted by a robot for reproduction. Many researchers handle this issue by using symbolic level representations coupled with primitive (atomic) actions [65, 120] or posture mapping with a motion capture system [55]. But this type of approach has clear drawbacks because it requires a large amount of prior knowledge to define the set of the primitive actions to generate robot motor control commands.

The generalization problem is related to the question of what to imitate from demonstrations, and it also has a dependency on the representation method to encode demonstrations. Hidden Markov Models (HMMs) or Gaussian Mixture Models (GMMs) are widely used for statistically learning a desired skill across multiple examples [17, 29, 86]. Here, each action/movement is encoded by a probability density function and analyzed with such statistically-based machine learning methods. However, it is still challenging to learn gen-

eralized actions, especially for accomplishing high-level complex tasks. Many researchers thus still use pre-defined primitive actions as prior knowledge, and then analyze the complex tasks based on the primitive actions. For example, Mülling et al. demonstrated that a robot can learn table tennis after decomposing a single complex task into multiple simple movement primitives, but it assumed that the robot already had all motor primitives for a striking motion to hit a ping pong ball [82].

One recent trend in LfD approaches is to consider a robot as an active learner that can ask questions to its teacher (a human user) when the robot is unsure what to do next during learning [15]. This learning strategy is particularly beneficial for Human-Robot Collaboration since it helps to understand each partner better and create a closer rapport. For instance, Tellex et al. presented an approach for enabling a robot to asking for help from the human partner when the robot fails to accomplish a given task [110]. The authors showed that the robot could recover from the failure better when the robot asked for more detailed help about their requests, such as “please hand me the black table leg” rather than just saying “help me.” A study by Nikolaidis and Shah also introduced an interactive training method called *Cross-training* that suggests switching the roles of human and robot during the training phase for improving human-robot teamwork [87]. They reported that a human-robot team’s performance was significantly improved by cross-training. Moreover, they showed that human participants perceived the robot more positively when they interactively switched their positions with their robot partner.

Although this line of work has many possibilities, it also has some limitations since it usually assumes that human teachers are physically located in the same place as the robot during the training phase. It thus requires much time and effort for the human partner to teach the robot even simple tasks. In addition, it is still challenging to apply the LfD paradigm for learning complex tasks that contains several individual motions. In

the remaining part of this chapter, we review recent approaches for learning human-robot collaborative tasks and then summarize this chapter with discussion.

## 2.2 LEARNING HUMAN-ROBOT COLLABORATIVE TASKS

The aim of this thesis is not only to build intelligent systems, but also to build “interactive” systems that can collaborate with humans. In order to build intelligent agents that would work together side by side with humans within the same workplaces, many human factors must be properly addressed. For example, researchers must ensure safety to prevent potential hazards by robots [61], and they also need to consider a human partner’s mental state (i.e., feelings, desires, intents, etc.) to make them feel more comfortable with robot co-workers [12, 38]. Moreover, since people perceive and react to robots differently [100], it is important to consider robot’s appearance and other factors if we want people treat a robot as their “work partner” or “friend.” The LfD paradigm cannot be used to for human-robot collaboration without considering the above human factors, despite the many other attractive points of LfD.

There has been a number of previous efforts to use the LfD paradigm for human-robot collaborative tasks while considering those human-centric issues [15, 27, 29, 85]. Although each paper handles human factors in different ways, we can consider the problems tackled by these approaches as a kind of uncertainty minimization problem, since most of the problems stem from unpredictable behaviors of humans/robots. Therefore, reducing the uncertainty in communication is a key concept of these research areas in the success of the learning process.

For example, many HRI researchers focus more on designing a way to help a human user easily understand their robot partner, rather than to develop techniques to just transfer knowledge from a human to a robot for replicating certain motions [12], since this

approach helps increase human’s predictability. Humans tend to feel uncomfortable when we are in unpredictable situations or when we are unable to understand someone’s intention or meaning, but we are comfortable as we become more familiar with the situations or each other. To be our partners, robots also need to understand human’s mental states, hence many researchers attempt to automatically detect social cues of people using various techniques [16, 78].

Another important perspective to employ the LfD paradigm for human-robot collaborative tasks is to make the learning process a bidirectional activity rather than passively observing human demonstrations [15]. This line of research is called *Interactive/Active Learning* and considers a robot as an active partner that provides feedback to the human teacher during the collaboration phase, and then uses the feedback to reduce uncertainty and makes learning new tasks more efficient [54].

Learning complex tasks is one of the most difficult remaining challenges in the field. It is considered as the ultimate goal of LfD-based approaches since we want to robots to automatically learn some high-level skills like “pick-up” instead of teaching them all the arm trajectories to accomplish the task. Although some research has been carried out on learning high-level actions [20, 34], there has been little discussion about how to learn high-level skills [80].

In this section, we begin by reviewing some recent approaches that attempt to build effective communication methods for learning human-robot collaborative tasks. We review research that employs *Interactive/Active Learning* methods in the following section.

### 2.2.1 INTENTION RECOGNITION

Much of the previous research on learning human-robot collaborative tasks has been carried out to recognize social cues such as body posture, facial expressions, direction of gaze, and



verbal cues in interactions, since they can give hints to a robot about the goal of the current task for learning. The robots are then able to use these hints to reduce the search spaces to learn the new task [9]. Various techniques are used to recognize a human user’s intentions like eye-gaze detection [16], speech recognition [78], and motion capture [4], but the most intuitive way to let a robot know about our (humans) thoughts or feelings is probably to use our own (natural) language.

Tellex et al. presented a new system for understanding human natural language to automatically generate robot control plans [111]. They introduced a new probabilistic graphical model called Generalized Grounding Graphs in order to transfer natural language control commands to corresponding groundings like target objects, places, and paths. The proposed approach is based on Conditional Random Fields (CRFs), which are one of the popular approaches in natural language processing. They trained the system on a new dataset collected from 45 subjects for 22 different videos using Amazon’s Mechanical Turk (AMT). This system was able to interpret high-level natural language control commands such as “put the tire pallet on the truck,” and generate control plans for the robot.

Recently, Misra et al. also introduced a new approach to interpret a user’s natural language instructions to generate robotic actions for manipulation tasks [78]. Their approach considered the ambiguity of natural language-based instructions caused by the fact that the same instructions can be interpreted differently according to context like the robot’s location and the state of the target objects. For example, an instruction such as “fill the cup with water” can be interpreted by the robot as either taking a cup and filling it with water from the tap, or taking a water bottle from the fridge first, and then pouring water into the cup. Misra et al. handled this ambiguity in instructions and the large variations given by human natural language using a CRF with an energy function that encodes natural language instructions into environment variables for manipulation tasks. This energy

function is composed of several nodes and factors that represent natural language (which is converted into a set of verb clauses), environment, and controller instruction. To train this model, they first created a new dataset, Verb-Environment-Instruction Library (VEIL)-300, which has six different tasks related to service robot scenarios like “making coffee” or “serving affogato.” The dataset contains natural language commands, environment information, and ground-truth instruction sequences that correspond to the commands. They trained their model on this dataset for mapping natural language commands to robot controller instructions. The accuracy of their model, 61.8%, outperformed all of their baselines on the validation set.

Motion capture is also widely used for making demonstration datasets for teaching robots in the field. Koenemann et al. presented a real-time motion capture system that enables a robot to imitate human whole-body motions [55]. In this approach, human motions were captured using inertial sensors attached to the human body with an Xsens MVN motion capture system. In order to reduce the computational cost, they used simplified human model that only considered the positions of the end-effectors (i.e., the position of the hands and feet) and the center of mass instead of considering a high number of parameters to represent all joint positions of the body. Then, they applied inverse kinematics to find joint angles given the positions of end-effectors, and generated robot motions while considering finding stable robot configurations instead of just focusing on imitating the human motions directly. They demonstrated their approach with a Nao humanoid robot, and showed the robot was able to imitate human whole-body motions with consideration for stabilization in real-time.

Gesture recognition is also extensively used for human-robot collaboration since gestures can be one of the effective communication channels between humans and robots for working together [70]. Various sensors (i.e., a depth camera and a wired glove) and algorithms

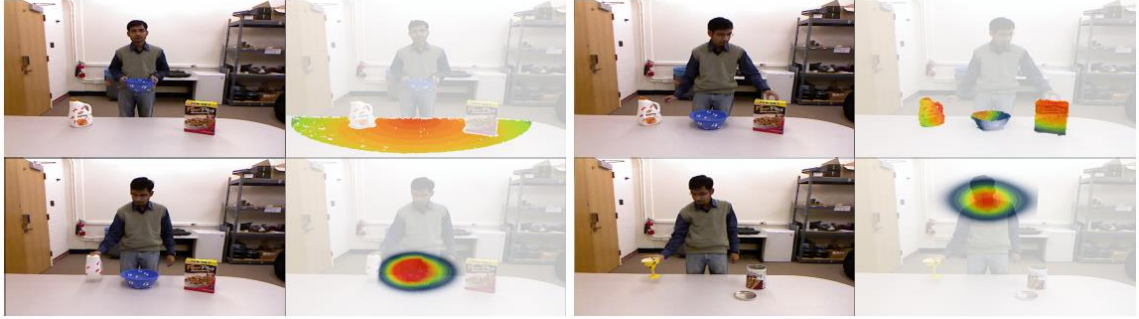


Figure 2.3: The learned affordance heat maps of four manipulation actions (place (top-left), reach (top-right), pour (bottom-left), and drink (bottom-right) in [57]. Here, red color represents the most likely object location according to the situation, while blue color indicates the most unlikely place. For example, the red color in heat map of top-left figure shows the most likely location to place the manipulation object on the table.

(i.e., Hidden Markov Models (HMMs)-based algorithms for modeling meaningful gestures and skeletal based algorithms for feature extraction and detection) are used in this line of research.

Mainprice and Berenson presented a new framework to recognize human’s intentions as early as possible [77]. The authors focused on building a framework for early detection of human motion in order to ensure safe robot behaviors when humans and robots are working together in close proximity. They modeled the human’s motion with a Gaussian Mixture Model (GMM) representation and performed Gaussian Mixture Regression (GMR) to predict the human’s future motion. The proposed approach then generated the robot motion given the prediction of human workspace occupancy obtained from the swept volume of predicted human motion trajectories. They demonstrated their approach on a PR2 robot simulation after training the framework on the human motion demonstrations of manipulation tasks on a table. They showed that the proposed approach was able to take into account the human motion predictions, so the robot could interact with the human co-worker more safely and efficiently in close proximity.

Another interesting keyword that can be used for understanding user’s intention is “*Affordance*.” The concept of affordance was defined by psychologist J. Gibson in 1966 for

describing inherent properties of an object that include all possible physical actions and transactions between the object and its environment [37]. Koppula and Saxena presented a CRF-based approach called an anticipatory temporal conditional random field (ATCRF) to predict future human activities based on object affordances [57]. Given the current observation of a human user’s pose and his/her surrounding environment, the goal of this approach is to anticipate what the user will do next. In order to achieve the goal, they first segmented an observed activity in time, then constructed a spatio-temporal graph based on the segmented sub-activities. The graph consists of four types of nodes (human pose, object affordance, object location, and sub-activity). They then augmented the constructed graph with anticipated nodes representing potential temporal segments. The authors demonstrated the proposed approach on the CAD-120 human activity dataset [58] and obtained 2.7% improvement on the state-of-the-art detection results in terms of success rate. They also reported that this approach achieved 75.4%, 69.2% and 58.1% accuracy in anticipating an activity 1, 3 and 10 seconds ahead of time, respectively.

Yi and Goodric presented a new framework for sharing information between robots and humans for task-oriented collaboration [121]. In this work, they considered a cordon and search mission, a kind of military tactic for searching out the enemy in an area, as a human-robot collaborative task that has to be solved by a human-robot team. Here, the authors assumed that a team supervisor (normally a human) assigns sub-tasks to his/her robot team members after decomposing the task, and then the robot team members are supposed to accomplish these given sub-tasks (i.e., searching a high risk sub-region for their human team members). They suggested the concept of a shared mental model for sharing knowledge about the current situation among all team members (robots and humans), so their framework was presented to help all human and robot team members understand each other correctly according to their task. Understanding all commands from natural language

is not easy, but the problem becomes easier in general if all team members know about the goal, so in this paper, the authors suggested to use a task-specific (oriented) grammar for converting a human supervisor’s verbal command into a sequence of way points.

### **2.2.2 LEGIBILITY OF ROBOT INTENTION**

Another direction of research in Human-Robot Collaboration (HRC) is designing more readable robot expressions and motions that convey clear intentions to their human partners. The objective of this line of research is finding effective ways to project a robot’s intention. Some people may think that this problem can easily be solved if a robot can speak like a human, but not all robots have a human like dialogue system. Moreover, we as human interpret of people’s thoughts or emotional states from other signs like facial expressions, voices, or even small body movements even when we do not understand other people’s speaking or we are not able to see whole-body motions of other people. The same kinds of signs are also very important and meaningful for natural Human-Robot Interaction, researchers have shown an increased interest in using non-verbal cues for human robot communication.

In 2005, Breazeal et al. showed that a robot’s non-verbal cues are important for building teamwork between humans and robots [13]. In this paper, they recruited 21 subjects and conducted a user study about task-oriented interactions with the robot Leonardo. Each subject first was asked to teach Leonardo the names of three different colored buttons (red, green and blue) which were located in front of the robot in its workspace, and then checked to see that the robot knew the names and locations of the buttons. After that, the subject was asked to guide the robot to turn on all of the buttons. After performing behavioral analysis of videos recorded during the experiment, they found that the robot’s non-verbal social cues (e.g., changes of gaze direction and eye blinks) helped humans read mental states

of the robot and improved human-robot task performance. The self-report results from the subjects also suggested that subjects perceived that the robot was more understandable when it showed non-verbal behaviors in addition to explicit expression.

Mutlu also showed how embodied cues like facial expressions, gaze, head gestures, arm gesture, social touch, social smile, etc. play important roles in communication [83]. He explored research on human communication, and found strong evidence about the hypothesis that embodied cues help achieve positive social and task outcomes in various domains of social interaction such as “Learning and Development” or “Motivation, Compliance, and Persuasion” in human-human communication. Thus, he finally suggested that HRI researchers should study the most effective ways to use such embodied cues for designing social robots, considering the relationship between particular embodied cues and outcomes.

His recent work with Sauppe is a good example of the importance of embodied cues for designing human collaborative robots [100]. In this paper, they studied how robots are treated by human co-workers in an industrial setting, while focusing on aspects of the robot’s design and context. The authors found that workers perceive the robots very differently according to various aspects like the physical appearance of the robots or their positions (roles) at their places of work. For example, workers who were supposed to operate the robot treated it as their “work partner” or “friend,” while maintenance and management staff just considered the robot as other industrial equipment. Another interesting finding is that human workers felt that the robots had some intelligence because of the robot’s eye movements, which suggested the robots knew what they were doing. Actually, they were pre-programmed movements: the robots just moved their eyes to follow the trajectory of their arms. However, even though those movements were simple, they helped the human workers understand the status of the robots and their next actions. Thus, it made human workers feel safe when they were working in close proximity to the robots, since they believed

the robots were able to convey their intentions through the eyes.

Takayama et al. applied animation principles to create readable robot behaviors [107]. In this paper, the authors created a robot animation which shows different robot behaviors according to their hypotheses (H1: showing forethought before performing an action would improve a robot’s readability, H2: showing a goal-oriented reaction to a task outcome would positively influence people’s subjective perceptions of the robot) and then measured how people described the robot’s intentions (before the action is performed) and how people perceived the robot in terms of some adjectives such as appealing and intelligent after conducting a video prototyping study with a total of 273 subjects. They found that people perceived the robot to be more appealing and their behaviors were more readable when the robot showed forethought before taking actions. They also discovered that showing a reaction made people feel that the robot was more intelligent. Even though this research is not LfD based, it shows potential benefits since the animation principles have been verified and successfully used to make a character by connecting its actions in animations. Furthermore, HRI researchers can design robot behaviors and test them using animation instead of building/programming of physical robot to test new robot motions.

It is worth noting that the readable robot behaviors are not always exactly the same as either the optimal behaviors of the robot to achieve its goal or expected robot behaviors that we can predict when we observe robot operations.

Dragan et al. focused on the difference between two types of robot motions, predictable robot motion and legible motion [27]. They argued that both robot motions are fundamentally different and often show contradictory properties. Here, predictable robot motions are those that match expected behaviors of observers (humans). On the other hand, legible robot motions are those that convey their intentions of behaviors clearly. In this research, the authors formalized legibility and predictability in the context of goal-directed robot

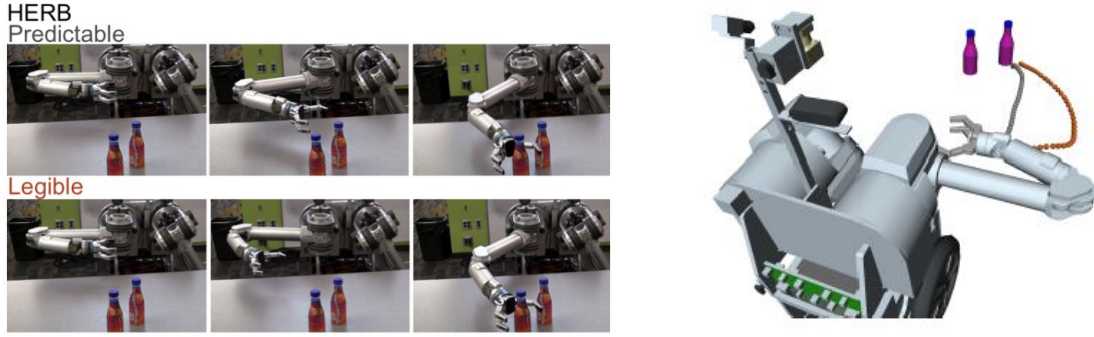


Figure 2.4: The difference between legibility and predictability of robot motion in [57]. The three sequential frames in the top-left show predictable robot motions which match the observer’s expectation, while the left-bottom row shows legible trajectories that convey the robot’s intention more clearly. The right figure shows the trajectories of the robot’s end effector. The gray shows predictable motions, while orange represents legible motions.

motions, then modeled both robot motions based on a cost optimization function which is designed in consideration of the principle of rational action. Finally, they demonstrated that two types of motions were contradictory through their experiments with three characters (a simulated point robot, the bi-manual robot mobile manipulator (HERB), and a human). They found that this difference between two properties derived from inferences in opposing directions, “action-to-goal” and “goal-to-action,” which refer to an observer’s ability to answer the questions: “what is the function of this action?” when she/he observes ongoing robot actions and “what action would achieve this goal?” when the observer knows the robot’s goal, respectively. Their experimental findings supported the theory in Psychology that humans interpret observed behaviors as goal-directed actions.

The same authors studied the effect of “*familiarization*” on the predictability of robot motion in their follow-up work [26]. This research originated from the idea of having users learn from robot demonstrations in order to increase their ability to predict robot motions (familiarization) because predictability is one of the keys for building collaborative robots that can work side by side with humans. This research direction is opposite from making robot motions more predictable, and gave us valuable insight about building more natural human-robot collaboration frameworks. They used the same methods that were used in



their previous work [26] to generate predictable robot motions, and then conducted a new user study to see the effect of familiarization on the robot motions. They recruited a total of 50 participants via AMT and conducted familiarization tests on two different types of robot motion (natural motion vs. unnatural motion), where the natural motion was defined being predictable without (or prior to) familiarization. In the experiment, each participant was asked to answer questions about the predictability of the robot motion before and after exposing them to examples of robot demonstration videos. They found that the robot motions became more predictable after familiarization, even though the familiarization was not enough for users to identify robot motions, especially when the robot operated in high-dimensional space with certain complex movements. The authors also reported that familiarization could help humans be more comfortable with robots and less natural robot motions hindered our ability to predict the motions.

Perlmutter et al. tried to make robots provide their internal states to human users to help them understand the robot’s thought and intentions, since we as humans are not able to judge what robots can see, hear, or infer in the same way that we can in human-human communication [92]. In this paper, the authors proposed visualization-based transparency mechanisms rather than developing a human-like verbal (or non-verbal) communication system for robots. The proposed visualization module is one kind of the tools that could be added on to a robotic perception system, which consists of three perception components (scene perception, pointing detection, and speech processing) to interpret the user’s commands. They conducted a user study with 20 participants with the proposed robotic system, and investigated the effect of their visualization-based transparency mechanisms. Their findings indicate that visualizations can help users communicate with the robot and understand its abilities even though some participants reported that they still prefer to have human-like transparency mechanisms with robots.

### 2.2.3 INTERACTIVE/ACTIVE LEARNING

In recent years, research on interactive/active learning has received considerable critical attention in the field. Cakmak and Thomaz introduced *Active Learning* to allow a robot to ask questions to its teacher (a human user) when the robot is unsure of what to do next during learning [14, 15]. In their article, they identified three types of queries (label, demonstration and feature queries) for an Active Learning based method in LfD and conducted two sets of experiments with human subjects. The first set of experiments was designed to investigate how humans ask questions in human-human collaboration scenarios with some levels of abstraction of the tasks, in consideration of employing similar scenarios for human-robot collaboration. The second set of experiments was designed to evaluate the use of the three types of queries in human-human collaboration scenarios. The authors found that participants perceived the robot as the smartest when it asked questions using feature queries, which directly ask about specific features like positions and rotations to manipulate target objects for learning a new task (e.g., “Should I keep this orientation at the start?”). They also reported that this type of query was the most commonly used in human learning (82%) even though it is the most challenging type of query for robots to produce automatically since it requires situational understanding. These findings provide guidelines to design good questions for building robots as an active learner in human-robot collaboration scenarios.

Tellex et al. presented an approach for a robot to receive help from its human partner when they work together for accomplishing a certain task [110]. They used a natural language generation system, called inverse semantics, for making a robot that can request help from the human partner in the form of natural language allowing the robot to recover from the failure based on their help. Since it is impossible to make a perfect robot that never fails, they focused on developing this recovery method based on a natural language



Figure 2.5: Active Learning allows a robot to ask questions to its teacher when the robot is unsure of what to do. This figure is obtained from the demonstration video of [15].

generation system for mapping from a desired human helping behavior that the robot would like the human to execute to words in natural language commands. This system was then used for generating requests when the robot needs assistance. When the robot detects failures using a motion capture system (VICON), their system first represents the failure in a simple symbolic language which indicates the desired human action, and then translates this symbolic representation to a natural language sentence using a context free grammar (CFG) to ask a human for assistance. In this research, the authors demonstrated their approach on a human-robot collaborative task of assembling a table, and then conducted a user study to evaluate the effectiveness of the proposed approach. The experimental results showed that it helped participants infer the requested action from the robot better than baselines approaches such as always using a general request (e.g., “Help me”) or generating requests using template based methods (e.g., “Hand me part 2”).

Knox et al. presented a case study of teaching a physically embodied robot by human feedback based on a framework called TAMER (Training an Agent Manually via Evaluative Reinforcement) that they previously proposed for robot learning from human reward [54]. In this paper, the authors focused on teaching interactive navigation behaviors to their

Mobile-Dexterous-Social (MDS) robot Nexi using human feedback as the only training resource. There were two buttons for providing positive or negative reward to the robot learner according to its state-action pair, and the robot then was able to be trained given the human reward. The authors taught a total of five navigation behaviors such as “Go to,” “Keep conversational distance,” and “Look away” to the robot, and then they tested the learned robot behaviors. However, they found that Nexi did not move properly after training due to issues of *transparency*. These transparency issues arose due to mismatches between the current state-action pair of the robot learner and what the human-trainer was observing. The authors pointed out that there were two main reasons for this confusion: 1) There can be a delay in the robot taking an action, creating mismatch between human observations and internal states of the robot, and 2) The perception system of the robot is not perfect, so the robot is not able to see some objects that the human trainer can. The authors suggested that researchers should address these transparency challenges when they employ a human feedback-based robot learning method for teaching a physically embodied robot.

Hausman et al. presented an approach based on the *interactive perception* paradigm which uses robot’s actuators for actively getting more information about the environment (world) when the robot is unsure about making a decision [43]. They proposed a particle filter-based approach to combine visual robotic perception with the outcomes of the robot’s manipulation actions in a probabilistic way, and the robot then found the best action to reduce uncertainty over articulated motion models given all sensory inputs. Here, the articulated motion models indicate the possible movements of objects such as certain directions (or rotations) of the objects that can be used for manipulating them. For example, a door of drawers or cabinets has parts that can be moved (also cannot be moved) for opening/closing it, which can provide useful information to a robot since it reduces the manipulation space.

In this work, they considered four types of articulated motion models: rigid, prismatic, rotational, and free-body, and then parametrized them with different numbers of variables according to the types. They demonstrated the proposed approach using a PR2 mobile manipulator. Their experimental results supported that the robot was able to effectively reduce uncertainty over models in four manipulation scenarios (opening and closing of a rotational cabinet door, moving a whiteboard eraser in a straight line, opening a locked drawer, and grasping a stapler on a table), and the robot then selected the best action based on a KL-divergence based information gain approach.

Nikolaidis and Shah introduced an interactive training method called *Cross-training* for improving human-robot teamwork [87]. A human and a robot are supposed to switch their roles during the training phase for learning a new collaborative task by cross-training. This training approach can be considered as a mutual adaptation process. They reported that a human-robot team’s performance significantly improved by cross-training for accomplishing a collaborative task, a simple place-and-drill, in their experimental results with human subjects. The authors also showed that participants who interactively switched their positions with their robot partner, Abbie, perceived the robot much more positively than the comparison group who trained with the robot using standard reinforcement learning methods in the post experimental survey. Their findings suggest that we are able to get better team performance with a robot partner for accomplishing certain tasks together when we switch our role with the robot during training phase in a way similar to human-human team training practices.

## 2.3 SUMMARY OF RELATED WORK

In this section, we reviewed some recent studies on robot learning from demonstration (LfD) while focusing on learning collaborative tasks. LfD is a very attractive research

direction for building a collaborative robot since it enables robots to automatically learn a new task from non-robotics experts (end-users). However, it is also challenging because of common theoretical issues like the *Correspondence Problem*, and the situation becomes more challenging when a robot needs to learn complex tasks. Moreover, researchers should also consider many human-centric issues such as safety, the human partner’s feelings, and intention. Since human-robot collaboration is not a task that a robot can perform alone (i.e., painting and assembly), but requires a robot to work side by side as a partner, researchers must consider human-centric issues.

Communication is an important challenge for human robot collaboration because communicating between a robot and a human can be very different from communicating between people, even though there are a lot of efforts to make this human-robot communication similar to human-human communication.

Several attempts have been made to use non-verbal cues such as facial expressions, gaze directions, and body gestures in human-robot communication since these signals can give additional useful information for natural and effective communication. In addition, some research has been carried out for designing human readable robot behaviors (predictable and legible robot motions) for conveying a robot’s intentions more clearly.

Another interesting line of research in the field employs an interactive/active based method for robot learning. These studies suggest viewing a robot as an active learner that can ask questions when the robot is unsure what is going on and what to do next. Furthermore, the robot can actively move itself for gathering more information for accomplishing/learning a new task in those kinds of situations.

One of the most difficult remaining challenges in the field is teaching complex tasks to a robot. People may want a robot to automatically find and learn all sub-tasks (i.e., picking-up a part, holding a part, and turning a screw) to accomplish a single complex task (i.e.,

assembling a table) instead of teaching all of them separately. However, this is technically challenging because there are no known approaches to find all necessary sub-tasks without any prior knowledge.

Recently, deep learning-based approaches have been widely used in many applications including object detection, scene segmentation, and learning robot motor control policy for grasping objects [67]. However, only a few previous studies have investigated applying deep learning for teaching robots human-collaborative tasks from demonstrations. In our view, the main reason is that it is hard to build the large-scale datasets, required for training deep learning methods. Different robots have different abilities with different body configurations and different people want to teach the robots different tasks, so all of them make the problem harder.

In this thesis, we propose to use hands as a medium to interpret human activities from human demonstration videos. One advantage of using hands to link humans and intelligent systems is that human hands play an integral role in physically interacting with the world. Moreover, hands might be one of the most frequently appearing objects in the first-person or action focused human demonstration videos which we suggest to use as training resources. Another advantage of focusing on human hands is that it reduces the correspondence problem for building interactive intelligent systems, since most robots have their own end effectors that can play the role of human hands.

However, the proposed approach may not solve all human-centric issues since imitating human behavior does not guarantee removing all ambiguity in communication between a human and a robot or solving safety issues. Furthermore, it is possible that a robot does not have any end effectors since each robot has its own unique appearance and functions. In such cases, we need to consider how to transfer the learned knowledge from human demonstrations to the intelligent agent. In spite of its limitations, we believe that our

approaches will open many possibilities for robot learning since we can easily collect enough training data for deep learning without a robot. Therefore, we suggest to learn human activities through the movements of human hands, and this work will generate fresh insight into the research of activity learning for intelligent systems.



## CHAPTER 3

### OBSERVING HUMAN BEHAVIORS IN DEMONSTRATION VIDEOS

In order to learn desired activities by watching human demonstration videos, we first build a perception system for observing the demonstrations. In this chapter, we describe our perception components to understand human behaviors according to different application scenarios. In particular, we focus on the movements of human hands in demonstration videos, since humans primarily use their hands for interacting with the physical world or communicating with others.

We first present a novel fully convolutional network to analyze human hand movements in human-human collaboration videos. We then extend the proposed approach for analyzing the state changes of objects in videos of daily activity. Finally, we investigate another new deep learning model for analyzing more fine-grain hand movements in videos of people playing a piano in the third section.

#### 3.1 ANALYZING HUMAN HANDS IN HUMAN-HUMAN INTERACTION VIDEOS

##### 3.1.1 INTRODUCTION

Our first application scenario is analyzing a camera wearer’s hand movements in first person perspective videos of two people collaborating. Given videos of the camera wearer executing collaborative behaviors for their partner, our objective is to learn such behaviors for making

the intelligent agents imitate the same behaviors for human-robot collaboration.

There has been previous work on robot activity learning from human videos [65,101,120], extending the previous concept of LfD. However, this work focused on learning grammar representations of human activities, modeling human activities as a sequence of motion primitives such as lift and poke. One of the drawbacks with these approaches is that they require labeled training data for recognizing a pre-defined set of primitive actions, and thus are limited in learning from scratch. Moreover, adding a new activity is another potential concern since it demands new labeled training data for each new action. In another approach, Koppula and Saxena suggested to directly learn object manipulation trajectories from human videos, however there are limits to how far the concept of object affordances can be taken for learning activities in human-robot collaboration scenarios since it assumed one-robot-one-object scenarios [56].

In this thesis, we introduce a new learning model using a fully convolutional network for future representation regression. We extend the state-of-the-art convolutional object detection network for the representation of human hands in training videos, and introduce the concept of using a fully convolutional network to regress the intermediate scene representation corresponding to a future frame. Combining these allows direct and explicit prediction of future hand locations (Figure 3.1) which then enables the robot to infer where to move its end-effectors to imitate the learned human behaviors. Our approach provides an important opportunity to learn human activities since it does not require activity labels, even though hand annotations are still required for initial training of the hand representation network. Furthermore, our networks are designed to function in real-time for the actual robot operation, and thus will make an important contribution to the field of robotics.

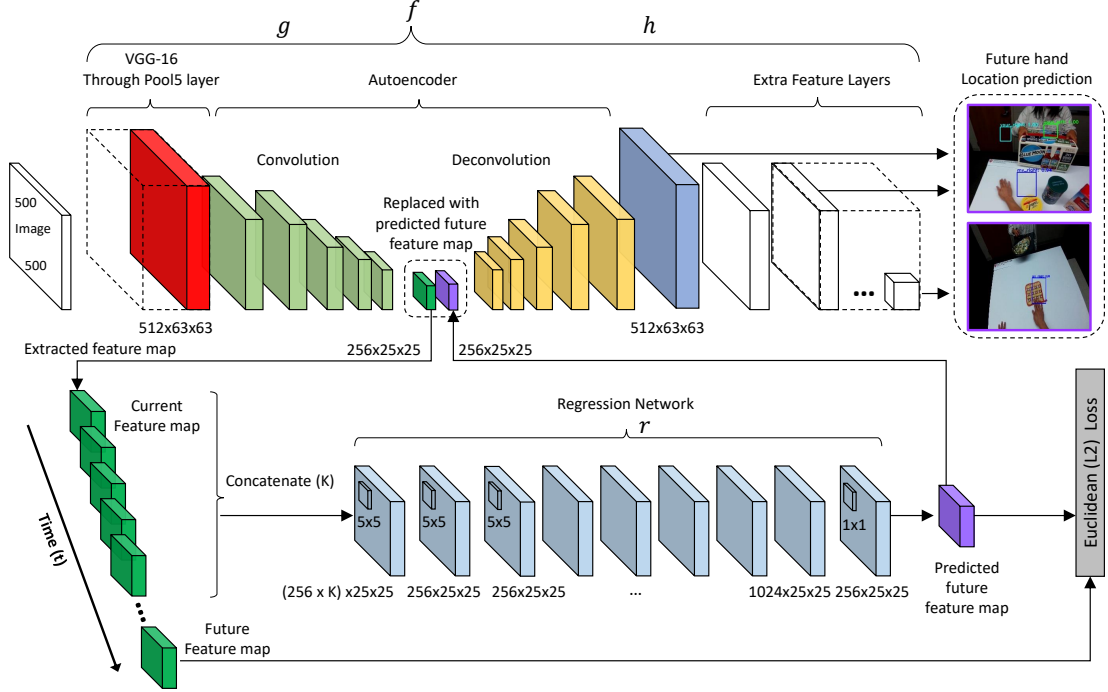


Figure 3.1: Overview of our perception component. Our perception component consists of two fully convolutional neural networks: The first network is an extended version of the state-of-the-art convolutional object detection network (SSD [71]) for representing human hands and estimating the bounding boxes (top). The second network is a future regression network to regress (i.e., predict) the intermediate scene representation corresponding to a future frame. This network does not require activity labels or hand/object labels in videos for its training.

### 3.1.2 APPROACH: CONVOLUTIONAL FUTURE REGRESSION

We employ two fully convolutional neural networks for analyzing the movements of hands. One network detects human hands in the videos and creates a hand-based scene representation. This is an extended version of the Single Shot MultiBox Detector (SSD) [71], which is one of the state-of-the-art object detectors. The second is a future regression network to model the change of this scene representation in future frames. Finally, the two neural networks are connected together and form a joint model for “future” hand detection given a current video frame.

The key idea of our approach is that we assume that intelligent systems also perceive the world from a first-person point of view, so they have a similar viewpoint to human first-person videos. Therefore, we can directly apply our perception component to either a

human or an intelligent system. This also allows the intelligent agent to predict what will happen next (i.e., where human hands and all interactive objects will be moved) by visually understanding the current situation.

### Hand Representation Network

Given a sequence of current frames, the goal of the hand representation network is to create a hand-based representation and estimate hand locations based on the representation. We construct this hand representation network by extending the SSD object detection framework, by inserting a fully convolutional auto-encoder having five convolutional layers followed by five deconvolutional layers for dimensionality reduction. This allows the approach to abstract an image with spatial information into a lower dimensional intermediate representation.

All our convolutional/deconvolutional layers use  $5 \times 5$  kernels and the number of filters for each convolutional layer are 512, 256, 128, 64, and 256, corresponding to the green convolutional layers in Figure. 3.1. After such convolutional layers, there are deconvolutional layers (yellow layers in Figure. 3.1), each having the symmetric number of filters: 256, 64, 128, 256, 512. We use stride 2 for the last convolutional layer for the dimensionality reduction and do not use pooling operation in any layer. We thus increase the number of filters from 64 to 256 to compensate for loss of spatial information at the last convolutional layer.

Let  $f$  denote the hand representation network given an image at time  $t$ . Then, this network can be considered as a combination of two sub functions,  $f = g \circ h$ :

$$\hat{\mathbf{Y}}_t = f(\hat{\mathbf{X}}_t) = h(\hat{\mathbf{F}}_t) = h(g(\hat{\mathbf{X}}_t)), \quad (3.1)$$

where a function  $g : \hat{\mathbf{X}} \rightarrow \hat{\mathbf{F}}$  denotes a feature extractor that takes an input video frame

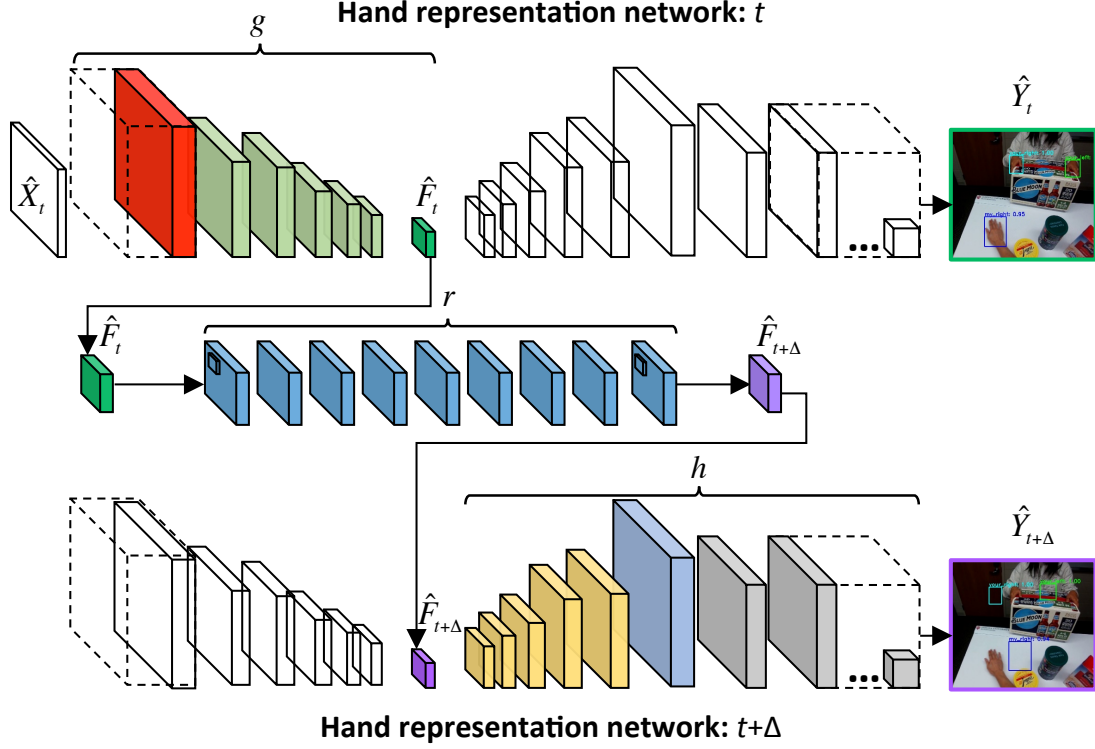


Figure 3.2: Data flow of our perception component for analyzing the movements of hands during the test phase. It enables predicting hands corresponding to the future frame. Only the colored layers are used for the prediction in the test phase.

and produces a compressed intermediate visual representation  $\hat{\mathbf{F}}$ , and  $h : \hat{\mathbf{F}} \rightarrow \hat{\mathbf{Y}}$  indicates a bounding box estimator which uses the compressed representation as input for locating hand boxes at time  $t$ . With the above formulation, the network can predict hand locations  $\hat{\mathbf{Y}}_t$  at time  $t$  after the training.

### Future Regression Network

Although the above hand representation network allows obtaining locations of hands in the ‘current’ frame, our objective is to get the ‘future’ hand locations  $\hat{\mathbf{Y}}_{t+\Delta}$  instead of their current locations  $\hat{\mathbf{Y}}_t$ .

We formulate this problem as a regression. The main idea is that the intermediate representation of the hand representation network  $\hat{\mathbf{F}}_t$  abstracts the hand-object information in the scene, and that we are able to take advantage of it to infer the future intermediate

representation  $\hat{\mathbf{F}}_{t+\Delta}$ . Once such regression becomes possible, we can simply plug-in the predicted future representation  $\hat{\mathbf{F}}_{t+\Delta}$  to the remaining part of the hand network (i.e.,  $h$ ) to obtain the final future hand prediction results. Therefore, we design a new network for predicting the intermediate scene representation corresponding to the future frame  $\hat{F}_{t+\Delta}$ , as a fully convolutional future regression network in Figure. 3.2.

Given a current scene representation  $\hat{\mathbf{F}}_t$  from the hand representation network, our future regression network ( $r$ ) predicts the future the intermediate scene representation  $\hat{\mathbf{F}}_{t+\Delta}$ :

$$\hat{\mathbf{F}}_{t+\Delta} = r_w(\hat{\mathbf{F}}_t). \quad (3.2)$$

It has seven convolutional layers having 256  $5 \times 5$  kernels. In addition, it has a layer with 1024  $13 \times 13$  kernels followed by the last layer that has 256  $1 \times 1$  kernels. We train the weights ( $w$ ) of the regression network with unlabeled first-person human activity videos using the following loss function:

$$\begin{aligned} w^* &= \arg \min_w \sum_{i,t} \|r_w(\hat{\mathbf{F}}_t^i) - \hat{\mathbf{F}}_{t+\Delta}^i\|_2^2 \\ &= \arg \min_w \sum_{i,t} \|r_w(g(\hat{\mathbf{X}}_t^i)) - \hat{\mathbf{F}}_{t+\Delta}^i\|_2^2 \end{aligned} \quad (3.3)$$

where  $\hat{\mathbf{X}}_t^i$  indicates a video frame at time  $t$  from video  $i$ , and  $\hat{\mathbf{F}}_t^i$  represents a feature map at time  $t$  from video  $i$ .

Our future regression network can use any intermediate scene representation from any intermediate layers of the hand network, but we use the one from auto-encoder for taking advantage of its lower dimensionality. Finally, the future scene representation  $\hat{\mathbf{F}}_{t+\Delta}$  is fed into the hand network for estimating hand bounding boxes corresponding to the future

frame to obtain future hand locations  $\hat{\mathbf{Y}}_{t+\Delta}$ ,

$$\hat{\mathbf{Y}}_{t+\Delta} = h(\hat{\mathbf{F}}_{t+\Delta}) \quad (3.4)$$

Figure 3.2 summarizes the data flow of our perception component during testing phase. Given a video frame  $\hat{\mathbf{X}}_t$  at time  $t$ , (1) we extract the intermediate scene representation  $\hat{\mathbf{F}}_t$  using the feature extractor ( $g$ ), and then (2) feed it into the future regression network ( $r$ ) to get the future scene representation  $\hat{\mathbf{F}}_{t+\Delta}$ . Next, (3) we feed  $\hat{\mathbf{F}}_{t+\Delta}$  into the box estimator ( $h$ ), and finally obtain the future estimated position of hands  $\hat{\mathbf{Y}}_{t+\Delta}$  at time  $t$ .

$$\hat{\mathbf{Y}}_{t+\Delta} = h(\hat{\mathbf{F}}_{t+\Delta}) = h(r(\hat{\mathbf{F}}_t)) = h(r(g(\hat{\mathbf{X}}_t))) \quad (3.5)$$

Furthermore, instead of using just a single frame (i.e., the current frame) for the future regression, we extend our network to take advantage of the previous  $K$  frames to obtain  $\hat{\mathbf{F}}_{t+\Delta}$  as illustrated in Figure.3.1:

$$\hat{\mathbf{Y}}_{t+\Delta} = h(r([g(\hat{\mathbf{X}}_t), \dots, g(\hat{\mathbf{X}}_{t-(K-1)})])). \quad (3.6)$$

The advantage of our formulation is that it allows us to capture temporal information of the movements of hands for predict future hand locations while considering the implicit activity and object context, even without explicit detection of objects in the scene. Our auto-encoder-based intermediate representation  $\hat{\mathbf{F}}_t^i$  abstracts the scene configuration by internally representing what objects/hands are currently in the scene and where they are, and our fully convolutional future regressor takes advantage of it for the prediction.

### 3.1.3 EXPERIMENTAL RESULTS

#### Datasets

We use two different types of datasets for training each neural network.

**EgoHands [6]:** This is a public dataset containing 48 first-person videos of people interacting in four types of activities (playing cards, playing chess, solving a puzzle, and playing Jenga). It has 4,800 frames with 15,053 ground-truth hand labels. Here, we added 466 frames with 1,267 ground-truth annotations to the original dataset to cover more hand postures. We use this dataset to learn our hand detection network, which is trained to locate hand boxes in a video frame.

**Unlabeled Human-Human Interaction Videos:** We created a new dataset with a total of 47 first-person videos of human-human collaboration scenarios, with each video clip ranging from 4 to 10 seconds. This is our main dataset for teaching a new task to our robot. It contains two types of tasks: (1) the camera wearer clearing all objects on a table while a partner (i.e., the other subject) approaches the table holding a heavy box to be placed on the table, and (2) the camera wearer pushing a trivet on the table toward to a partner while he/she approaches holding a hot cooking pan. These are unlabeled videos without any activity or hand annotations, and we trained our convolutional regression network using this dataset.

#### Baselines

We quantitatively compared our proposed future hand prediction network with three different baselines. (i) **Hand-crafted representation** uses a manually-crafted state representation based on explicit object and hand detection. It encodes relative distances between all interactive objects in our two scenarios, and uses them to predict future hand location using neural network-based regression. More specifically, this baseline detects objects using



Method	Evaluation		
	Precision	Recall	F-measure
Hand-crafted representation	$0.30 \pm 0.37$	$0.15 \pm 0.19$	$0.20 \pm 0.25$
Hands only	$4.78 \pm 3.70$	$5.06 \pm 4.06$	$4.87 \pm 3.81$
SSD with future Annotations	$27.53 \pm 23.36$	$9.09 \pm 8.96$	$13.23 \pm 12.62$
Deep Regressor (ours): K=1	$27.04 \pm 16.50$	$21.71 \pm 14.71$	$23.45 \pm 14.99$
Deep Regressor (ours): K=5	$29.97 \pm 15.37$	$23.89 \pm 16.45$	$25.40 \pm 15.51$
Deep Regressor (ours): K=10	<b><math>36.58 \pm 16.91</math></b>	<b><math>28.78 \pm 17.96</math></b>	<b><math>30.90 \pm 17.02</math></b>

Table 3.1: Evaluation of future hand locations prediction (1 sec later) with baselines, in terms of precision, recall, and f-measure.

KAZE features [2] and hands using the CNN based hand detector in [6], and then computes relative distances between all objects and hands for building the state representation, which is a 20 dimensional vector. Then, we train a new neural network which has five fully connected layers on the same human-human interaction videos to predict future hand locations based on the 20 dimensional feature vectors. **(ii) Hands only** uses frame-based hand detection results for future regression. It predicts future hand locations solely based on the hand detection results of the current frame. We extract hand locations from all frames of the interaction videos using our hand representation network, and then train another seven-layer neural network for future hand location prediction. **(iii) SSD with future annotations** uses the original SSD model [71] trained on the EgoHands dataset. Instead of training the model to infer the current hand locations given the input frame, we fine-tuned this model on the EgoHands dataset with “future” locations of hands as the ground truth labels. We also added 466 new frames to the original EgoHands dataset for making this baseline since the original dataset has some repetitive patterns of hand movements.

### Evaluation of our future hand prediction

We first evaluate our future hand prediction network in terms of precision, recall, and F-measure, and compared them against the above baselines. In the first evaluation, we

Method	Mean Pixel Distance
Hand-crafted representation	143.85 $\pm$ 48.77
Hands only	247.88 $\pm$ 121.94
SSD with future Annotations	58.58 $\pm$ 36.76
Deep Regressor (ours): K=1	51.31 $\pm$ 39.10
Deep Regressor (ours): K=5	51.41 $\pm$ 38.46
Deep Regressor (ours): K=10	<b>46.66</b> $\pm$ 36.92

Table 3.2: Mean pixel distance between ground truth and predictions. The video resolution was 1280x960.

Method	Mean Pixel Distance
Hand-crafted representation	121.48 $\pm$ 87.36
Hands only	264.52 $\pm$ 148.15
SSD with future Annotations	48.63 $\pm$ 39.04
Deep Regressor (ours): K=1	40.08 $\pm$ 32.72
Deep Regressor (ours): K=5	40.46 $\pm$ 39.52
Deep Regressor (ours): K=10	<b>36.78</b> $\pm$ 36.70

Table 3.3: Mean pixel distance between ground truth and predicted position of right hand. In this experiment, we only consider the camera wearer’s right hand predictions since we attempt to learn collaborative behaviors from the camera wearer’s reactions to the interaction situation. The video resolution was 1280x960.

made our approach predict bounding boxes of human hands in the future frame given the current image frame. We measured the “intersection over union” ratio between areas of each predicted box and ground truth (future) hand locations. Only when the ratio was greater than 0.5 was the predicted box accepted as a true positive. In this experiment, we randomly split the set of our Human-Human Interaction Videos into the training and testing sets, so 32 videos were used for training sets and remaining 15 videos were used for testing sets in a total of 47 videos.

Table 3.1 shows quantitative results of our future hand prediction. Here,  $K$  represents the number of frames we used as input for our regression network. Our  $\Delta$  was 30 frames to forecast hand locations one second ahead of time. We observe that our approach significantly outperforms all baselines, including state-of-the-art SSD modified for hand prediction. Our proposed network with  $K = 10$  yielded the best performance in terms of all three metrics,

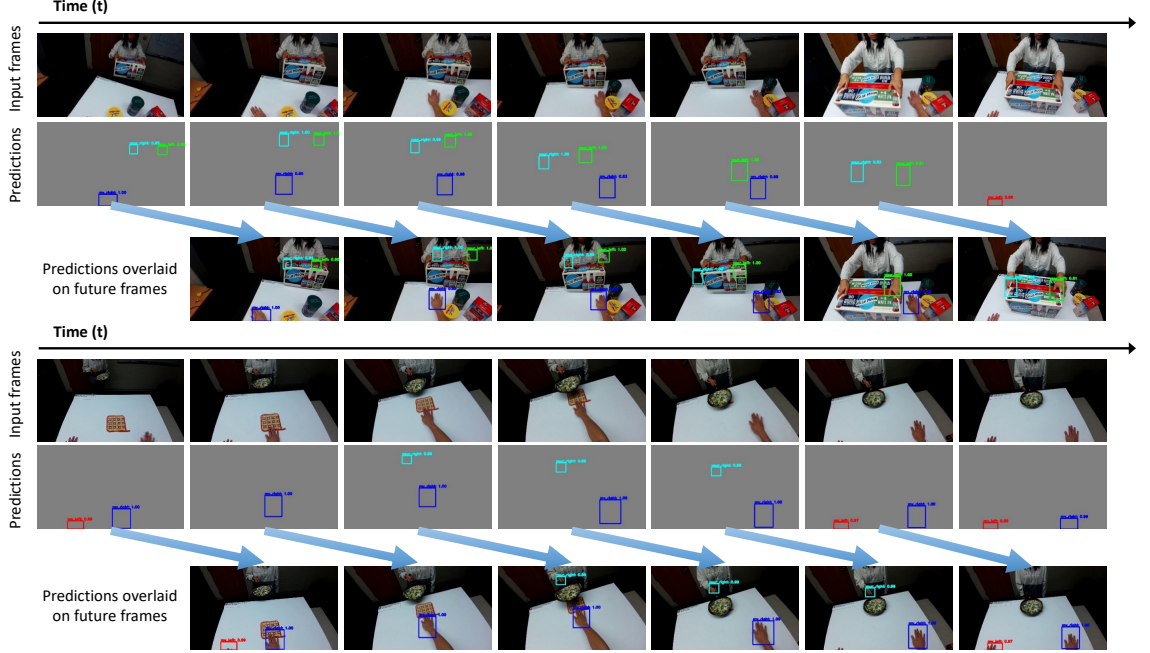


Figure 3.3: Two examples of our visual prediction. The first example is the activity of clearing the table, and the second example is the activity of pushing the trivet toward the person holding a cooking pan. The first row shows the input frames and the second row shows our future hand prediction results. In the third row, we overlaid our predictions on “future” frames. Red boxes correspond to the predicted ‘my left hand’ locations, blue boxes correspond to ‘my right hand’, green boxes correspond to the opponent’s left hand, and the cyan boxes correspond to the opponent’s right hand. The frames were captured every one second.

with about 30.9 score in F-measure.

In our second evaluation, we measured mean pixel distance between ground truth locations and the predicted positions of hands. We measured this only when both the ground truth and the predictions are present in the same frame. Table 3.2 shows the mean pixel distance errors for all four types of hands (my left, my right, your left, and your right). Once more, the results confirm that our approaches outperform the performance of the baselines.

We also compared mean pixel distances of these methods while only considering the camera wearer’s right hand predictions, since position of the right hand is more important for learning collaborative behaviors in our experimental scenarios. Table 3.3 shows mean pixel distance between ground truth and predicted position of ‘my right hand,’ we can see that performance of our approaches is superior to all the baselines. Examples of our visual

predictions results are illustrated in Figure. 3.3.

## 3.2 ANALYZING OBJECTS IN VIDEOS OF DAILY LIVING

### 3.2.1 INTRODUCTION

Our second application scenario is learning the state changes of objects (i.e, changes in location of the objects and which objects will appear/disappear in the future) from first-person videos of daily activities such as brushing teeth and watching television. This is also a necessary ability for intelligent systems that work together with people, since collaborative tasks require understanding how the objects near them will be moved by human partners.

Recently, a considerable literature has grown up around the theme of forecasting a future event in computer vision. For example, Park et al. investigated predicting future trajectories of ego-motion and tried to discover space occluded by a foreground object from egocentric stereo images [89]. In 2017, Ma et al. reported a new predictive models to forecast future behaviors of multiple pedestrians using game theory and deep learning-based visual analysis [76]. In a study conducted by Luo et al, an unsupervised learning approach was presented to predict a scene representation that encodes long-term motion dependencies as a sequence of basic 3D motion flows for activity recognition [75]. Even though they did not demonstrate their approach for recognizing events in future frames, it showed the possibility to predict future human actions based on the proposed scene representation.

However, research to date has not yet targeted forecasting explicit locations of objects appearing in videos. Vondrick et al. investigated forecasting presence of objects, but they did not attempt to predict their locations [114]. Some work has attempted to predict future video frames directly [35, 72]; so far, however, there has been little discussion about forecasting object location in future frames and previous approaches also assumed the objects to be already present in the scene.

In this thesis, we extend our previous proposed perception network for hand detection to take advantage of motion-domain features for predicting “future” object locations given a current video frame. We predict object locations in the future frame even when they are not visible in a sequence of current image frames. We believe this is the first attempt to present a method to explicitly forecast location of objects in future frames using a fully convolutional network.

### **3.2.2 APPROACH: TWO-STREAM NETWORK WITH A TEMPORAL STREAM**

The objective of this study is to predict object location in a future scene given a sequence of current image frames. Here, we propose a new two-stream convolutional neural network architecture, by extending our previous proposed perception network for hand detection which we introduced in the previous section. This design was inspired by the success of two-stream networks for activity recognition [103]. We use X and Y gradients of optical flow as input to a temporal stream network to capture temporal motion patterns in human daily activity videos, while the spatial stream network receives an image frame as input. The same base network, VGG-16 [104], is used to extract both the spatial and temporal information from the two different inputs, and then two feature maps are combined together after feature extraction. To be specific, we combine two feature maps using a fusion layer with a  $1 \times 1$  kernel before the auto-encoder component of the proposed network, which constructs a single combined  $256 \times 25 \times 25$  feature blob which contains both spatial and temporal information of the input video. This data fusion approach can be considered as the early fusion in the published paper by Feichtenhofer et al. [33]. The main advantage of applying this early-fusion method instead of late-fusion is that we can reduce the amount of computations in our temporal stream to stack optical flows from multiple frames, since

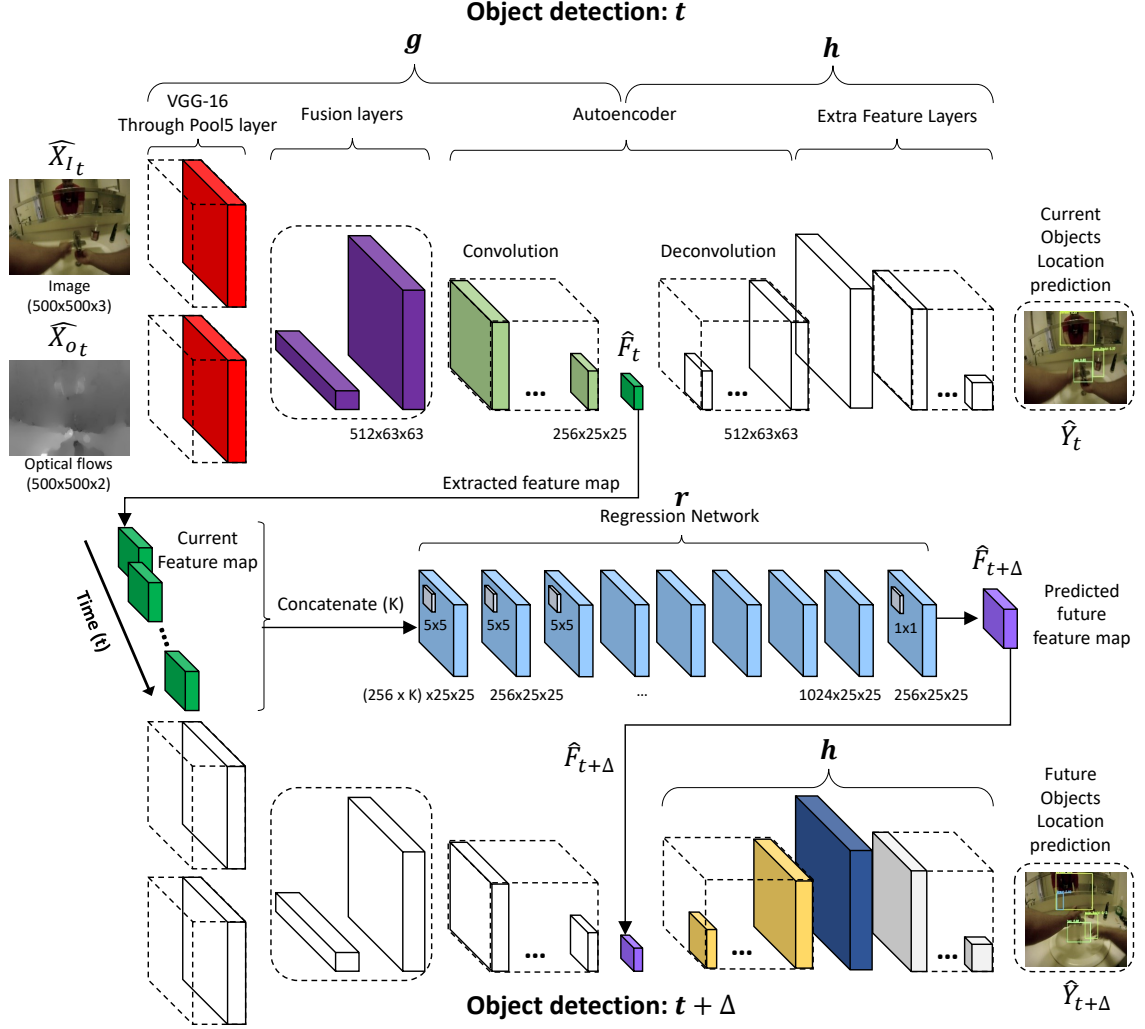


Figure 3.4: Overview of our two-stream network, consisting of two fully convolutional neural networks. The first is the two-stream object detection network (1st row and 3rd row of the figure). The 1st row and the 3rd row are duplicates of the same model. The second network is a fully convolutional regression network used to predict the future intermediate scene representation (the 2nd row). Only the colored layers are used in the actual testing stage.

our future regression network can use this combined scene representation which already has temporal information in the form of the feature vector.

Figure. 3.4 illustrates our two-stream network for analyzing the state changes of objects in videos of daily living. The spatial stream takes the image frame as an input ( $\hat{\mathbf{X}}_{\mathbf{I}t}$ ), while the temporal stream receives the corresponding X and Y gradients of optical flow ( $\hat{\mathbf{X}}_{\mathbf{O}t}$ ). Thus, the original  $f$  of our hand/object representation network in Equation 3.1 is changed

to take two input variables  $\hat{\mathbf{X}}_{\mathbf{I}t}$  and  $\hat{\mathbf{X}}_{\mathbf{O}t}$ .

$$\hat{\mathbf{Y}}_t = f(\hat{\mathbf{X}}_{\mathbf{I}t}, \hat{\mathbf{X}}_{\mathbf{O}t}) = h(\hat{\mathbf{F}}_t) = h(g(\hat{\mathbf{X}}_{\mathbf{I}t}, \hat{\mathbf{X}}_{\mathbf{O}t})), \quad (3.7)$$

Note that  $\hat{\mathbf{X}}_{\mathbf{O}t}$  is calculated from image  $I_{t-1}$  and  $I_t$ , so no future information after time  $t$  is used and the other functions ( $g$ , and  $h$ ) and the scene representation ( $\hat{\mathbf{F}}_t$ ) are the same as in Equation 3.1. The objective function for future regression to predict the future scene representation ( $\hat{\mathbf{F}}_{t+\Delta}$ ) is also changed to take both the inputs, where  $r$  denotes our future regression network.

$$\begin{aligned} w^* &= \arg \min_w \sum_{i,t} \|r_w(\hat{\mathbf{F}}_t^i) - \hat{\mathbf{F}}_{t+\Delta}^i\|_2^2 \\ &= \arg \min_w \sum_{i,t} \|r_w(g(\hat{\mathbf{X}}_{\mathbf{I}t}^i, \hat{\mathbf{X}}_{\mathbf{O}t}^i)) - g(\hat{\mathbf{X}}_{\mathbf{I}t+\Delta}^i, \hat{\mathbf{X}}_{\mathbf{O}t+\Delta}^i)\|_2^2 \end{aligned} \quad (3.8)$$

Finally, the future scene representation is fed to the function  $h$  to forecast object locations ( $\hat{\mathbf{Y}}_{t+\Delta}$ ) corresponding to the future frame, in our previous work in Section 3.1

$$\hat{\mathbf{Y}}_{t+\Delta} = h(\hat{\mathbf{F}}_{t+\Delta}) = h(r(\hat{\mathbf{F}}_t)) = h(r(g(\hat{\mathbf{X}}_{\mathbf{I}t}, \hat{\mathbf{X}}_{\mathbf{O}t}))). \quad (3.9)$$

### 3.2.3 EXPERIMENTAL RESULTS

#### Datasets

**Activities of Daily Living (ADL)** [94]: We use this public dataset that contains 20 first-person videos of 18 daily activities, such as making tea and doing laundry, for analyzing the state changes of objects related to human activities. The dataset is challenging because frames display a significant amount of motion blur caused by the camera wearer’s movement and the annotations are noisy. There are 43 types of objects in the dataset which are

Forecast	Method	dish	door	utensil	cup	oven	person	soap	tap	tbrush	tpaste	towel	trashc	tv	remote	mAP
5 secs later	Vondrick [114]	4.1	22.2	5.7	16.4	17.5	8.4	19.5	20.6	9.2	5.3	5.6	4.2	8.0	2.6	10.7
	SSD with future annotation	18.9	17.6	0.0	28.1	7.1	23.0	0.0	37.7	0.0	0.0	0.0	0.0	20.4	0.0	10.9
	SSD (two-stream)	13.5	22.4	0.0	15.2	4.1	14.3	39.8	21.4	0.0	0.0	0.0	0.4	48.4	0.0	12.8
	Ours (one-stream K=1)	34.4	37.0	18.9	19.2	24.3	75.1	70.0	55.0	23.8	6.7	16.6	2.1	57.5	61.7	35.9
	Ours (one-stream K=10)	35.1	42.4	22.2	29.9	37.9	69.9	68.0	67.6	21.7	47.7	17.7	5.2	30.5	36.4	38.0
	Ours (two-stream K=1)	38.2	44.1	23.8	29.1	37.2	73.1	67.1	60.6	12.2	38.0	13.7	4.4	37.2	58.5	38.4
	Ours (two-stream K=10)	35.7	44.0	24.2	29.3	39.6	75.7	68.9	63.2	20.4	47.2	18.2	4.6	40.4	60.3	<b>40.8</b>

Table 3.4: Future object *presence* forecast (5 seconds later) evaluation on the ADL dataset.

annotated with ground truth labels and bounding boxes. Among them, we trained our model (and the baselines) for the 15 most common categories to compare with the baseline [114]. We split the ADL dataset into four sets, and trained our object perception model on three sets and used the remaining set for evaluation.

## Baselines

We compared our two-stream network with three different baselines: **(i) SSD with future annotations** is the same baseline that was used in the evaluation of our hand detection network (described in Section 3.1) but trained on the ADL dataset with “future” ground truth object locations. This enables the model to directly regress future object locations given the current frame. **(ii) SSD (two-stream)** uses the same training scheme as the first baseline, but extended to have two-stream inputs to see how much the temporal stream helps to improve the performance of forecasting the state changes of objects. **(iii) Ours (one-stream K=10)** is our approach for hands described in Section 3.1, but trained on the ADL dataset for future object location prediction. It thus only has the spatial stream without the temporal stream.



Forecast	Method	dish	door	utensil	cup	oven	person	soap	tap	tbrush	tpaste	towel	trashc	tv	remote	mAP
1 sec later	SSD with future annotation	0.0	0.5	0.0	0.0	0.0	0.2	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	SSD (two-stream)	1.6	12.4	0.0	0.9	5.1	9.6	0.0	2.8	0.0	0.0	0.0	0.0	29.8	1.3	4.5
	Ours (one-stream K=1)	0.4	15.0	1.2	2.6	13.8	43.4	4.4	19.0	0.0	0.0	0.3	0.0	16.0	18.8	9.6
	Ours (one-stream K=10)	0.4	14.0	0.0	0.7	16.1	45.8	5.4	22.9	0.0	0.0	0.9	0.0	20.5	6.8	9.5
	Ours (two-stream K=1)	7.3	19.6	1.9	1.8	37.2	26.2	11.6	33.8	0.0	1.4	1.5	0.8	11.0	11.9	11.9
	Ours (two-stream K=10)	3.8	10.1	1.8	5.5	19.0	59.6	2.8	41.8	0.0	0.0	3.4	0.0	15.9	45.2	<b>14.9</b>
5 secs later	SSD with future annotation	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.2
	SSD (two-stream)	2.0	11.7	0.0	3.2	10.2	0.5	3.2	0.0	0.0	0.0	0.0	0.0	20.0	0.0	3.7
	Ours (one-stream K=1)	0.5	10.8	0.0	0.3	16.5	10.4	3.2	8.2	0.0	0.0	0.7	0.0	3.9	1.7	4.0
	Ours (one-stream K=10)	0.2	10.7	0.0	0.2	0.7	35.7	1.3	5.6	0.0	0.0	0.5	0.0	3.8	1.2	4.7
	Ours (two-stream K=1)	1.5	9.8	0.4	0.4	24.1	17.0	8.6	15.8	0.0	0.0	1.6	0.2	7.5	5.8	6.6
	Ours (two-stream K=10)	0.7	4.7	0.0	5.0	9.7	35.6	0.7	10.5	0.0	0.0	1.4	0.0	15.0	24.8	<b>7.7</b>

Table 3.5: Future object *location* forecast evaluation using the ADL dataset.

### Object presence forecast

We first evaluated our approach to predict ‘presence’ of objects in future frames. To be specific, we estimated whether the objects will exist in the future frame or not, regardless location, given a sequence of current video frames. Here, we trained our model and all three baselines to predict presence of objects 5 seconds ahead of time.

Table 3.4 shows our result with the baselines. We observe that our two-stream approach outperforms all baselines in terms of mAP. It is also worth noting that our approach without any future frame information (K=1) significantly outperforms the results reported in [114] by more than 20 mAP. Furthermore, our K=10 model shows the best performance over all baselines, thus confirming that motion information helps object presence prediction as well.

### Object location forecast

We next measured the object location forecast accuracy. Similar to hand location prediction in Section 3.1, we estimated future bounding box locations (1 or 5 seconds later) and compared with the same baselines that were used in the previous evaluation. We report the accuracy of our prediction with the baselines in Table 3.5 in terms of average precision of

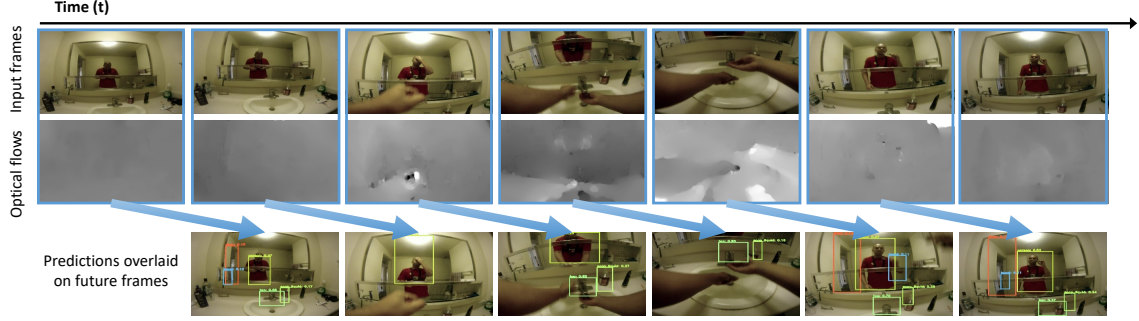


Figure 3.5: Example object location forecasts on the ADL dataset. The first two rows show the input frames and corresponding optical flows that are fed to our two-stream network. In the bottom row, we overlaid our predictions of object locations on “future” frames.

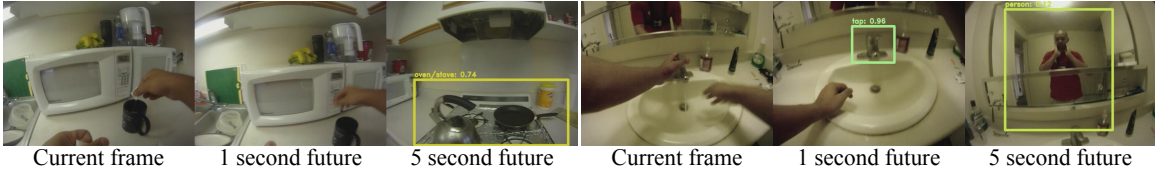


Figure 3.6: Example object location forecasts in two different time setting (1 second and 5 second later). Here, we overlaid predicted future bounding boxes on the actual “future” frames again. Our approach correctly predicted the locations of three objects (oven, tap, and person) by observing a sequence of current frames.

each object category.

As Table 3.5 shows, the overall accuracy of all methods was poor due to the many challenges of the ADL dataset. For example, objects often unpredictably appear/disappear in future frames despite that they exist in the current video frame. Nevertheless, we observe that our approach significantly outperforms all the SSD baselines including our one-stream approach. Figure 3.6 shows some examples of our object location forecasts.

### Hand location forecast

Additionally, we evaluated the performance of our approach to predict future hand locations using our unlabeled human interaction dataset (used in the previous section) to confirm the benefit of the temporal-stream for predicting future hand locations.

Table 3.6 shows quantitative results of 1-second future hand prediction on the human interaction dataset. The proposed two-stream models performed better than our one-stream

Method	Evaluation		
	Precision	Recall	F-measure
SSD with future Annotations	27.53 $\pm$ 23.36	9.09 $\pm$ 8.96	13.23 $\pm$ 12.62
Ours (one-stream K=10)	36.58 $\pm$ 16.91	28.78 $\pm$ 17.96	30.90 $\pm$ 17.02
Ours (two-stream K=1)	37.21 $\pm$ 22.49	26.69 $\pm$ 14.28	30.21 $\pm$ 16.07
Ours (two-stream K=5)	37.41 $\pm$ 22.97	26.19 $\pm$ 14.93	30.06 $\pm$ 17.16
Ours (two-stream K=10)	<b>42.89</b> $\pm$ 23.61	<b>30.46</b> $\pm$ 13.08	<b>34.18</b> $\pm$ 16.48

Table 3.6: Evaluation of future hand locations prediction (1 sec later) with two-stream network on Human Interaction dataset.

model, indicating the temporal stream is helpful to predict future locations. Our proposed model with K= 10 yields the best performance in terms of precision, recall, and f-measure.

### 3.3 ANALYZING PRESSED NOTES AND FINGER MOVEMENTS OF PEOPLE PLAYING PIANO

#### 3.3.1 INTRODUCTION

Our third application scenario is observing people playing a piano. An objective of this research is to develop an interactive piano tutoring system that provides real-time feedback to learners about their play like pressed keys, hand movements, posture, etc. In particular, we have two primary aims: 1) to determine which notes on a piano are being played at any moment in time, 2) to identify which finger is pressing each note.

In recent years, there has been an increasing interest in building intelligent music tutoring systems that apply Artificial Intelligence (AI) technology in music education [7, 46, 90]. A wide range of technologies has been employed for music teaching and learning, from Music Information Retrieval (MIR) techniques [24] to Augmented Reality (AR) with AI technology [21, 69].

Traditionally, researchers attempted to employ rule-based expert systems that store teaching materials, answers, and comments for evaluating a learner’s performance and giv-

ing feedback [22, 46]. MIDI (Musical Instrument Digital Interface) based digital musical instruments were commonly used for avoiding the problem of music transcription — converting audio signals of an acoustic instrument into a structured audio format like a music score sheet or MIDI file. However, traditional approaches are limited in providing real-time feedback which is important for a learner to check their mistake as early as possible. Moreover, there is a lot of demand to develop intelligent music tutoring systems that allow users to play acoustic instruments to learn a new instrument.

Recently, considerable literature has grown around Automatic Music Transcription (AMT) research, which aims to solve the music transcription problem by automatically generating the structured musical representation from an acoustic signal [8, 113]. AMT is a fundamental problem for a wide range of music applications, such as music information retrieval and interactive music systems, and various technologies have been attempted to detect pitch, rhythm, and onset (offset) in musical signals. Furthermore, a number of authors have considered building real-time processing systems that not only handle pre-recorded audio, but also live streaming of music [23, 25]. So far, however, very little attention has been paid to using the other sensory inputs like a video stream in addition to audio signals.

In this thesis, we introduce a novel two-stream convolutional neural network that takes video and audio inputs together for detecting pressed notes and finger movements of people playing piano. Although some research has been carried out on employing computer vision approaches [1, 108] on this problem, there have been few publications on music analysis with audio-visual fusion [91, 124]. In addition, previous published studies are difficult to extend since they require a carefully engineered event detection pipeline.

We formulate our two problems (note detection and finger identification) as object detection with multi-task learning rather than standard image classification. This view is especially useful for analyzing the piano player’s performance since it reduces the search

space for detecting pressed notes and identify fingers. We extend the Single Shot MultiBox Detector (SSD) [71] to take corresponding audio signals and an image frame to resolve ambiguities caused by finger or key occlusions, and design the model to focus on a single octave and hand for reducing the searching space. We report experiments measuring recognition accuracy and demonstrate that our approach is able to detect pressed piano keys and the piano player’s fingers with high accuracy.

### 3.3.2 RELATED WORK

#### **Intelligent Musical Instrument Tutoring**

There is a growing body of literature that applies Artificial Intelligence (AI) technology to building intelligent tutoring systems for learning various musical instruments such as guitar [7], piano [22] and violin [122]. The purpose of a intelligent musical instrument tutoring system is to help students learn how to play musical instruments by providing proper feedback after listening and analyzing the student’s playing. Much of the current literature on intelligent music tutoring systems pay particular attention to audio signal processing for analyzing the user’s performance [24, 90], which is a natural direction since the system is developed for music. However, there are limitations, since these systems often requires specifically designed instruments and controllers [24] for avoiding the problem of music transcription. Recent developments in Automatic Music Transcription (AMT) open the possibility that users can play acoustic instruments with the tutoring system [8, 113]. Nevertheless, the fundamental problem still remains since it only can give feedback about rhythm and sound. For example, hand posture or the position of fingers are also very important for learning to play piano; so far, however, very little attention has been paid to providing feedback about the student’s posture.

## Computer Vision in Music Analysis

In this thesis, we apply computer vision techniques for analyzing the pressed piano keys and the pianist’s finger movements in videos of people playing piano. Computer vision can play an important role in providing proper feedback about a student’s fingering and hand position on the piano. It also can help resolve ambiguities in the audio signals caused by complex interacting harmonics. There has been previous work on applying computer vision techniques for music analysis. Akbari *et al.* created a four-stage image processing pipeline based on Hough line transforms [47] for pressed piano keys detection [1]. Takegawa *et al.* attached color markers to the pianist’s finger nails, and then applied a simple color-based image processing pipeline with some musical rules for analyzing the pianist’s fingering movements [108]. Johnson *et al.* used a depth camera with Histograms of Oriented Gradients (HOG) features for detecting pianist hand posture [50]. However, there have been few attempts at integrating the computer vision approaches with audio signals to complement the limitations of each feature. Although a few studies have investigated multimodal fusion for music analysis [91, 124], their approaches are hard to expand to other musical instruments due to hardware requirements [124] and specific engineering design of the systems [91].

## Deep Learning in Music Analysis

In recent years, deep learning has emerged as a powerful tool for many AI applications from object detection [71, 96] to learning motor control policies for robotic applications [66]. It also has become popular in Music Information Retrieval (MIR) research, and many researchers have applied deep learning for various applications such as automatic music transcriptions of drum [113], piano [44, 102], chord detection [126], and music recommendation [68]. Most studies in the field of MIR, however, have only focused on audio signals, and only a few deal with multimodal fusion for music analysis [88].

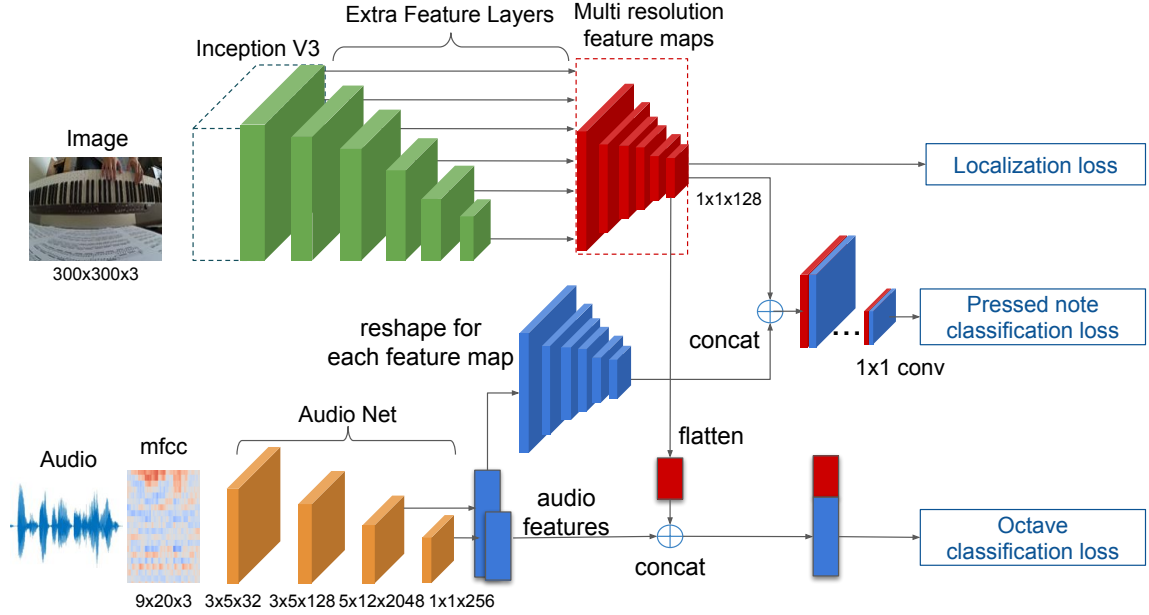


Figure 3.7: Outline of our two-stream architecture. The top row is the original SSD model with a different base network to handle visual stream input. We simply replace VGG16 [104] with Inception V3 [106] to forward more elaborate feature for constructing multi resolution feature maps. The bottom row is a four-layer CNN to handle audio stream. We employ MFCC for audio feature extraction, and take a late fusion approach to integrate the audio and visual feature vectors. Since the audio features do not have the same spatial information as the visual features, we concatenate them along the depth axis for each multi-resolution feature map after reshaping the audio features and do not use the audio features to compute localization loss. Our model is designed to focus on the piano key movements within a single octave to reduce the search space from 88 keys to 12 keys.

### 3.3.3 APPROACH: MULTI-TASK LEARNING WITH VIDEO-AUDIO FUSION

The objective of this study is to detect pressed piano notes and identify which fingers pressed each note at any moment in time by observing people playing a piano. In this section, we describe our two-stream audio-visual fusion network that learns to detect pressed notes and fingers from videos of people playing a piano.

#### Pressed Piano Notes Detection

We can formulate the first problem as image classification, which is the task of assigning to each video frame a label corresponding to pressed notes from 88 piano keys. We then use

a state-of-the art image classification network [127], and train on a dataset which contains people playing a piano with ground truth key pressed labels. The benefit of this approach is that we can achieve reasonable performance without any specific modification. However, there are certain drawbacks associated with the use of this formulation. One of these is that piano keys visually look very similar, so it is challenging to distinguish different pressed piano keys from all 88 keys. The other limitation of this formulation is that it does not exploit the other major sources of evidence like audio signals and the pianist’s hand movements, both of which are related to the pressed piano notes.

## Architecture

In this thesis, we formulate this problem as multi-task learning with audio-visual data fusion instead of simply adapting this standard image classification model. Our model focuses on the movements of piano keys in a single octave which contains 12 keys (7 white keys and 5 blacks), and use audio signals corresponding to the current image frame for boosting the performance of the classifier. The main ideas behinds our approach are: (1) piano keys of same notes at different octaves visually look exactly the same; the only difference between them are their locations on the piano, (2) audio signals help resolve ambiguity caused by finger or key occlusions, and (3) visual features also help resolve ambiguities in the audio signals caused by complex interacting harmonics.

Figure 3.7 shows the overall architecture of our model for analyzing pianist accuracy and form. We extend the state-of-the-art convolutional object detection network (SSD [71]) for multi-task learning, with an additional audio-stream to handle audio signals of video. Here, we define three tasks to analyze piano playing: 1) localization to find octave sections on the piano, 2) pressed piano note classification to identify the pressed piano keys in a single octave, and 3) octave classification to identify which octaves are played at any given



moment. Our model takes two inputs: the current image frame of people playing a piano and the audio feature map which represents the audio signals corresponding to the image frame.

To be specific, this audio feature map is constructed from 20-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) features for 100 milliseconds segments of video (which correspond 10 consecutive image frames of the input video, our video is recorded in 60 frames per second). We obtain 9 temporal feature sets with 100 ms for the window size, and then compute the first and second order derivatives of the MFCC features for making they have three channels like an RGB image. Each constructed audio feature map thus finally has a dimensional of  $9 \times 20 \times 3$ . The input videos are recorded with an image resolution of  $1920 \times 1080$  at 60Hz using a camera placed camera over the piano, but we resize the original image frames to  $300 \times 300$  in the preprocessing phase for feeding them to our model.

To integrate both visual and audio features, we take a late fusion approach which concatenates two feature vectors immediately before the final score functions. For pressed piano note classification, we extract audio features from the 3rd and 4th layers of the audio net, and then concatenate audio features along the depth axis of the multi-resolution image feature maps since the proposed model separately predicts the confidences per each default box. Once audio-visual data are concatenated, we employ  $1 \times 1$  convolution to take advantage of each feature vector for making a final decision. We also employ audio-visual data fusion for octave classification. Since octave classification is not related to locations of bounding boxes, we only use the last feature map from multi resolution feature maps of each data stream to predict one octave category at a time. For localization, we do not use the audio features since audio features do not have the spatial information needed to predict the location of one octave section on piano.

## Training

We extend the original objective function in SSD for handling multi-task learning. The extended objective function consists of three loss functions: (1) localization loss ( $L_{loc}$ ), (2) pressed piano note classification loss ( $L_{cls\_note}$ ), and (3) octave classification loss ( $L_{cls\_octave}$ ). The overall objective function is a weighted sum of these losses:

$$L(x, y, c_{note}, c_{octave}, l, g) = \frac{1}{N} (L_{cls\_note}(x, c_{note}) + \alpha L_{loc}(x, l, g)) + \beta L_{cls\_octave}(y, c_{octave}) \quad (3.10)$$

where  $N$  is the number of matched bounding boxes,  $x$  is a binary indicator (0 or 1) for matching the default box to the ground truth box of the piano note category  $p$ ,  $y$  is a binary indicator for matching the input image frame with the ground truth octave classification label of category  $q$ ,  $c_{note}$  and  $c_{octave}$  indicate confidence scores of pressed piano note classification in single octave and octave classification respectively, and  $l$  and  $g$  represent the locations of predicted box and the ground truth box.

For pressed piano note classification loss, we use a sigmoid function instead of the softmax that is used in the original SSD for multi-class, multi-label classification. Thus, the confidence loss is changed to:

$$L_{cls\_note}(x, c_{note}) = - \sum_{i \in Pos}^N x_{ij}^p \log(\hat{c}_i^p) - \sum_{i \in Neg} \log(\hat{c}_i^0) \text{ where } \hat{c}_i^p = \frac{1}{1 + \exp(-c_i^p)} \quad (3.11)$$

Here,  $i$  and  $j$  represent the box number ( $i$ -th and  $j$ -th) of the default box and the ground truth box respectively.

Likewise, we use a sigmoid function with cross-entropy loss for octave classification:

$$L_{cls\_octave}(y, c_{octave}) = -y^q \log(\hat{c}^q) - (1 - y^q) \log(1 - \hat{c}^q) \text{ where } \hat{c}^q = \frac{1}{1 + \exp(-c^q)} \quad (3.12)$$

We set the weight terms  $\alpha$  and  $\beta$  to 1 by cross validation and use the original localization loss, which is a Smooth L1 loss to regress location parameters of the predicted bounding boxes.

$$\begin{aligned}
L_{loc}(x, l, g) &= \sum_{i \in Pos}^N \sum_{m \in \{cx, cy, w, h\}} x_{ij}^k \text{smooth}_{L1}(l_i^m - \hat{g}_i^m) \\
\hat{g}_j^{cx} &= (g_j^{cx} - d_i^{cx})/d_i^w & \hat{g}_j^{cy} &= (g_j^{cy} - d_i^{cy})/d_i^h \\
\hat{g}_j^w &= \log\left(\frac{g_j^w}{d_i^w}\right) & \hat{g}_j^h &= \log\left(\frac{g_j^h}{d_i^h}\right)
\end{aligned} \tag{3.13}$$

Here,  $(cx, cy)$  indicates the offsets for the center of the default bounding box( $d$ ), and  $w$  and  $h$  represent width and height of the bounding box.

## Identifying Fingers Used to Press Notes

### Architecture

For identifying which fingers are used to press a note on a piano, we frame the problem as object detection and employ the same architecture without octave classification or the audio stream since the audio signals do not contribute to distinguishing different fingers. This problem is more challenging because the piano player’s fingers move very fast and they are small. Furthermore, the same finger looks differently depending on hand posture, and is often occluded by other fingers.

We propose to use the key pressed information obtained by the first network for reducing the search space to detect fingers. We assume that input videos are recorded with similar camera angles, and then use the key pressed information to crop the input image frames based on the rough locations of the pressed key on the piano. For example, we can remove the very left and right sides of the input image if the middle C is given by the first network. In this thesis, we crop out about 30% of the original input images as a preprocessing phase,

and then feed them to the network to identify used fingers.

## **Training**

One problem of the object detection formulation is that it requires more expensive annotations because the network needs bounding boxes of the target objects in training. In order to reduce this annotation cost, we use a public dataset for hand detection in images [6] for obtaining bounding boxes of fingers. Since human fingers are jointed digits on our hands, we adopt these hand bounding boxes for fingers with some offsets to ensure they contain all five fingers. Therefore, we first train our network on the EgoHands dataset [6], and then apply the trained model to construct our own dataset for finger identification. We assume that hands are located nearby in adjacent image frames, and thus use the previous bounding box locations for the current frame if the network trained on the EgoHands dataset fails to detect hands in a frame. Finally, we create our own dataset of finger detection by integrating these bounding boxes with finger numbers which are obtained according to the musical scores, and then train the proposed network on our new introduced dataset.

### **3.3.4 EXPERIMENTAL RESULTS**

We conducted two sets of experiments to evaluate the proposed architecture and compare it to various baselines. In the first set of experiments, we focus on testing the accuracy of our model for pressed piano notes detection. In the second set of experiments, we evaluate the accuracy of our approach for identifying fingers used to press notes.

## **Datasets**

We created a new dataset for observing video of people playing piano and learning models for detecting pressed keys and fingers from the video. Figure 3.8 shows the pipeline that we used for generating our dataset given three different input files: videos, MIDI files, and

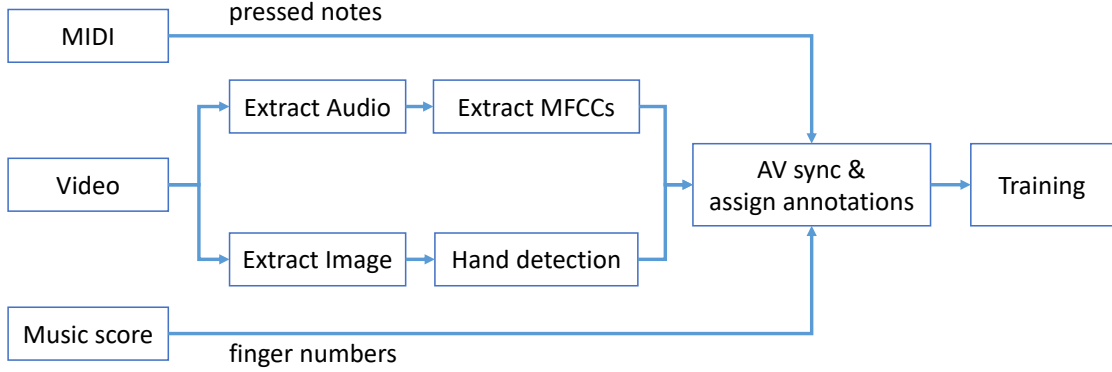


Figure 3.8: The pipeline to create our dataset of Hanon Exercises.

music scores. First, we extracted image frames and audio from the input video, and then applied the pre-trained hand detector on image frames to obtain bounding boxes of fingers. For the audio stream, we extracted MFCC features with 100 ms window size from the audio, and constructed MFCC-based audio images with the first and second order derivatives of the extracted MFCC features for representing the audio signals. In order to make ground truth labels of pressed notes and used fingers, we recorded MIDI files while a person was playing the piano and then used them with music scores for annotating our dataset. Finally, we synchronized the image frames and corresponding audio images with the annotations of finger numbers, bounding boxes, and pressed notes. Note that 6 consecutive image frames are synchronized with the same audio image since the videos in our dataset were recorded of 60 fps.

### Hanon Exercises

We used the Hanon Exercises [42], which are widely used for piano teachers and students to strengthen hands and fingers and build basic techniques for the music to create our dataset. Hanon Exercises might not be fun music that many people like to play, but we chose these since they evenly cover many piano notes with repeated playing patterns, which



Figure 3.9: Our piano room with an experimental setup and an example of Hanon Exercises that we used for constructing our own datasets. We recorded MIDI files while a person was playing the piano and then used them with music scores for annotating our dataset.

thus yields naturally balanced and directed data for the pressed notes detection. They are also beneficial for finger identification for a similar reason since the Hanon Exercises have all finger numbers on the score and they are designed to exercise all five fingers evenly. Therefore, we can easily create a large-scale balanced dataset by automatically assigning the pressed note and finger annotations based on MIDI events and the repeated finger patterns with the scores. Figure 3.9 shows our piano room for the experimental setup and the first few bars of Hanon Exercises number 1.

In this thesis, we prepared two datasets of people playing the Hanon Exercises. First, we collected videos of a person playing a piano with only one hand to create a relatively easier dataset, “One Hand Hanon.” We then constructed more advanced level dataset that contains both hands playing the same exercises, and named it as “Two Hands Hanon.”

**One Hand Hanon:** This dataset contains a total of 10 videos of a person playing Hanon exercises 1 to 5 with one hand, and each video clip ranges from 50 to 120 seconds. A player played each exercise twice using different hands (left and right) for recording these videos. In total, we collected 35,332 frames with ground-truth annotations. We split this dataset into five sets according to the exercise number, trained our model on exercises 1 to 3 (23,555

Method	Accuracy
<b>Using a Single Sensory Input:</b>	
Video Only (Inception V3 [106])	56.43%
Audio Only (Audio Net)	41.10%
<b>Video and Audio Data Fusion:</b>	
Two-stream w/o Multi-Task (Inception V3 + Audio Net)	75.05%
<b>Multi-Task Learning to focus on a Single Octave:</b>	
Video Only w/ Multi-Task (Inception V3 + Focusing a Single Octave)	82.37%
Two-stream w/ Multi-Task (Ours, Inception V3 + Audio Net + Focusing a Single Octave)	<b>85.69%</b>

Table 3.7: Pressed Notes Detection Accuracy on One Hand Hanon dataset.

frames), and then used the remaining exercises 4 and 5 (11,777 frames) for evaluation.

**Two Hands Hanon:** This dataset contains a total of 5 videos of a person playing the same Hanon exercises 1 to 5 with both hands, and each video clip ranges from 50 to 240 seconds. In total, we collected 51,596 frames with ground-truth annotations. Similar to One Hand Hanon Exercises, we split this dataset into five sets with regard to the exercise number, and trained our model on exercises 2 to 4 (36,115 frames) and then the remaining exercises 1 and 5 (15,481 frames) were used for evaluation. It is worth noting that this is a multi-label dataset for octave classification since Hanon Exercises have two octaves that are supposed to be played at the same time when played them with both hands.

## Evaluation

### Pressed Notes Detection

We first evaluated the accuracy of the proposed architecture for pressed notes detection. We compared our video-audio fusion model based on multi-task formulation with four different baselines. (i) **Video Only** is a baseline that only uses video frames as input to the classifier to identify the pressed piano keys. It thus formulates the problem as a standard image classification problem, and we used Inception V3 [106] for this baseline. Similar to this first baseline, (ii) **Audio Only** baseline also uses a single sensory input, but it uses audio signals

Method	Accuracy
<b>Using a Single Sensory Input:</b>	
Video Only (Inception V3 [106])	46.33%
Audio Only (Audio Net)	39.63%
<b>Video and Audio Data Fusion:</b>	
Two-stream w/o Multi-Task (Inception V3 + Audio Net)	65.33%
<b>Multi-Task Learning to focus on a Single Octave:</b>	
Video Only w/ Multi-Task (Inception V3 + Focusing a Single Octave)	65.82%
Two-stream w/ Multi-Task (Ours, Inception V3 + Audio Net + Focusing a Single Octave)	<b>75.37%</b>

Table 3.8: Pressed Notes Detection Accuracy on Two Hands Hanon dataset.

instead of the image frames. We used our Audio Net which is a four layer CNN described in Figure 3.7 for this baseline. **(iii) Two-stream w/o Multi-Task** uses audio-visual data fusion without our multi-task formulation which is designed to focus on the key movements in a single octave. This baseline uses Inception V3 and the Audio Net to handle each sensory input respectively, and then takes a late-fusion approach for integrating both inputs. **(iv) Video Only w/ Multi-Task** is a baseline that uses our multi-task formulation, but it only takes a video stream for detecting the pressed notes.

We measured classification accuracy for this first evaluation, which is a percentage of correctly classified images given total image frames of the test set. Since our approach and the Video Only w/ Multi-Task baseline can produce more than one output at a time with different bounding boxes, we picked the single predicted box that had the highest confidence score in each image, and then assigned its predicted class (pressed note) to the image for both approaches. In addition, we only accepted the image as a true positive when the image was correctly classified by each approach for both the octave and pressed piano notes classifications. We trained our model using RMSProp [45] for 50k iterations with learning rate  $10^{-4}$ , 0.9 momentum, 0.9 decay, and batch size 32 on both Hanon datasets for this experiment.

Table 3.7 shows the pressed notes detection accuracy of the proposed approach with



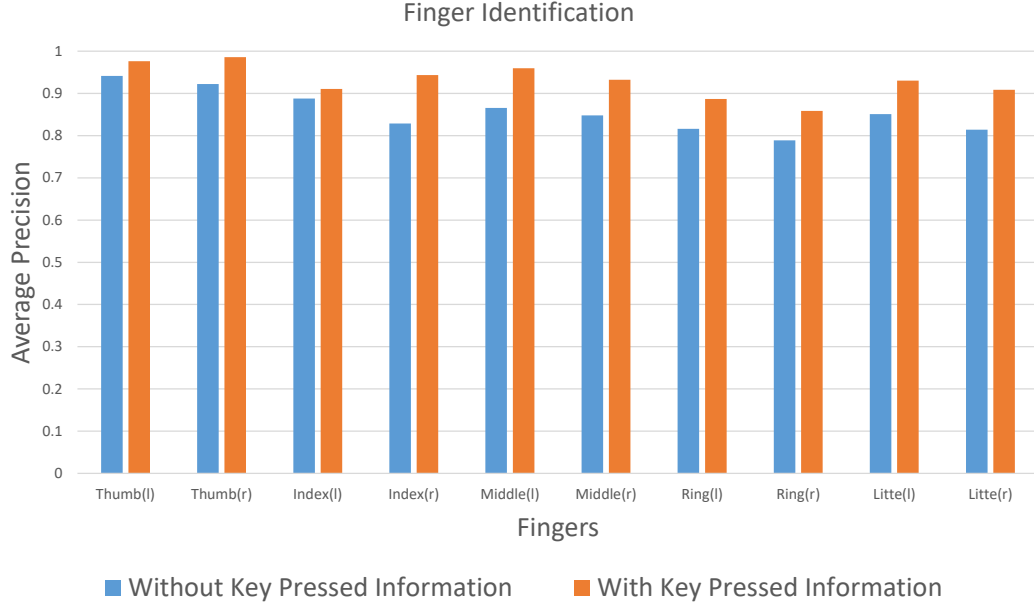


Figure 3.10: The accuracy of the proposed approach to identify fingers used to press piano notes. We evaluated the performance of two approaches (with key pressed information vs. without key pressed information) in terms of average precision. The x-axis shows finger names and (l) and (r) indicate the left and right hand respectively.

our four baselines on the One Hand Hanon dataset. We observe that our two-stream approach with multi-task learning formulation outperforms all baselines in terms of accuracy. Our model yielded 85.69 % pressed notes detection accuracy, and the experimental result confirms that our multi-task formulation and additional audio stream are able to boost the performance of the classifier.

We next measured the classification accuracy on the Two Hands Hanon dataset. We used the same baselines, training strategies, and hyperparameters for this experiment. However, we replaced the final softmax function of all baseline approaches with the sigmoid function since the Two Hands dataset changes the problem to a multi-label classification (because it contains two labels for each image). For our approach, we picked the top two predicted boxes based on the confidence scores, then assigned each image their predicted classes.

Table 3.8 shows the pressed notes detection results on this Two Hands Hanon dataset.

Once more, the results confirm that our approaches outperform the performance of the baselines in terms of classification accuracy.

### **Used Fingers Identification**

Our second set of experiments evaluated the accuracy of identifying fingers used to press notes on a piano. The purpose of these experiments was to validate our object detection formulation with the processing pipeline that was used to generate the finger identification dataset and to check how much the key pressed information helps for identifying fingers. For these experiments, we trained our object detector on the One Hand Hanon for detecting fingers, with and without a pre-processing stage to crop the input image frame based on key pressed information. We then measured used finger detection accuracy in terms of average precision.

Figure 3.10 illustrates used finger detection accuracy of the two approaches. From the figure, we can clearly observe that pressed notes information is beneficial to detect used fingers to press piano keys. The network achieved better accuracy for all fingers in terms of average precision when it used key pressed information. The model with the pre-processing step yielded 0.929 for mean average precision (mAP), which was more accurate than the model without key pressed information (0.856 in mAP).

## CHAPTER 4

### BUILDING INTERACTIVE SYSTEMS FOR NATURAL INTERACTION WITH PEOPLE

The specific purpose of this thesis is to build intelligent and interactive systems that can understand people’s behaviors and interact with them. Therefore, our activity learning should extend beyond perceiving human behaviors and consider a knowledge transfer problem, to transfer learned activities from observations to an intelligent system. Moreover, many human-centric issues should be addressed properly for building “interactive” systems. For example, researchers must handle safety issues to prevent potential hazards to humans working with intelligent agents [61], and also take the human partner’s mental states (i.e., feelings, desires, intents, etc.) into account for natural human-agent collaboration [12].

In this chapter, we present two different approaches for transferring knowledge from the proposed perception component (which we described in Section 3.1) to two different types of robots. First, we show a relatively simple method which directly maps the robot control commands to each human hand gesture. We use a small low-cost drone, Parrot AR Drone 2.0, as a hardware platform for demonstrating the approach to generate drone control commands and discuss its limitations. After that we introduce our new approach to imitate observed human actions with a humanoid robot, Baxter, which has two seven degree-of-freedom arms. Finally, we conduct a user study with the Baxter robot to evaluate the proposed approach and explore designable factors for making a robot more acceptable in social scenarios based on a focus group interview involving the drone.

## 4.1 FROM PERCEPTION TO INTELLIGENT AGENTS

A considerable amount of literature has been published on learning robot control policies based on visual input through a camera [36,66,93,125]. Recent trends in this line of research have employed convolutional neural networks (CNNs) to learn the robot control policies end-to-end for directly mapping raw image observations to the robot control commands, and they have shown promising results for many robotics applications, such as object grasping [36,93], manipulation [66,125], and autonomous driving [19].

However, the main disadvantage with this kind of approach is that it often requires expensive hardware equipment, like a specific type of robot for obtaining training samples, and demands a huge amount of time [67,93] to collect those examples. Furthermore, we must handle the Correspondence Problem to learn activities from human demonstrations, so we still need a bridge to link the perception part of the robot and the robot control policies.

For this reason, much of the previous research on human activity learning for robotics applications has employed grammar representations to model human activities for generating robot actions according to its camera input [65,118,120]. Since this allows a robot to encode and decode human activities based on the grammar, the robot can learn a new task from human demonstrations and replicate them. However, there are certain problems with the use of grammar representations. One of these is that they require us to manually define all possible atomic actions for generating the robot control commands. Another drawback is that they require separate control functions for each atomic action and the level of the control functions can vary according to each action.

In this thesis, we first employ a relatively simple approach that directly maps the robot control commands to human hand gestures for human-drone interaction. This can be a good example of the use of a robot for gathering training samples to minimize the knowledge

transfer cost from its perception part. We then introduce a new manipulation network to regress the target joint angles of the robot arms to generate robot control commands for imitating human collaborative behaviors.

## **4.2 FORECASTING HAND GESTURES FOR HUMAN-DRONE INTERACTION**

### **4.2.1 INTRODUCTION**

Our first application scenario is building an early recognition-based control system for human-drone interaction. The main objective of this research is to identify human hand gestures one second before they are actually completed, based on our perception component described in Section 3.1, and generate drone control commands to respond to the hand gestures. The hypothesis behinds this approach is that early recognition of human activities may help build more natural human-robot interaction, since it enables robots to respond quickly to a human partner.

In this application scenario, we introduce a dataset of humans interacting with a drone (Parrot AR.Drone 2.0) through gestures, and use it to train our activity learning approach. We consider two interaction scenarios with different interaction distances: (1) drone delivery, in which the person directs a drone to deliver a small object, and (2) self portrait, in which the person directs the drone to take a photo. We define five hand gestures (selfie, stop, come, go, change altitude) for the human to use to command the drone, since gestures are a natural interaction modality [18]. Finally, we experimentally confirm that our approach enables a drone to forecast future human gestures and generates the drone control commands according to the gestures.

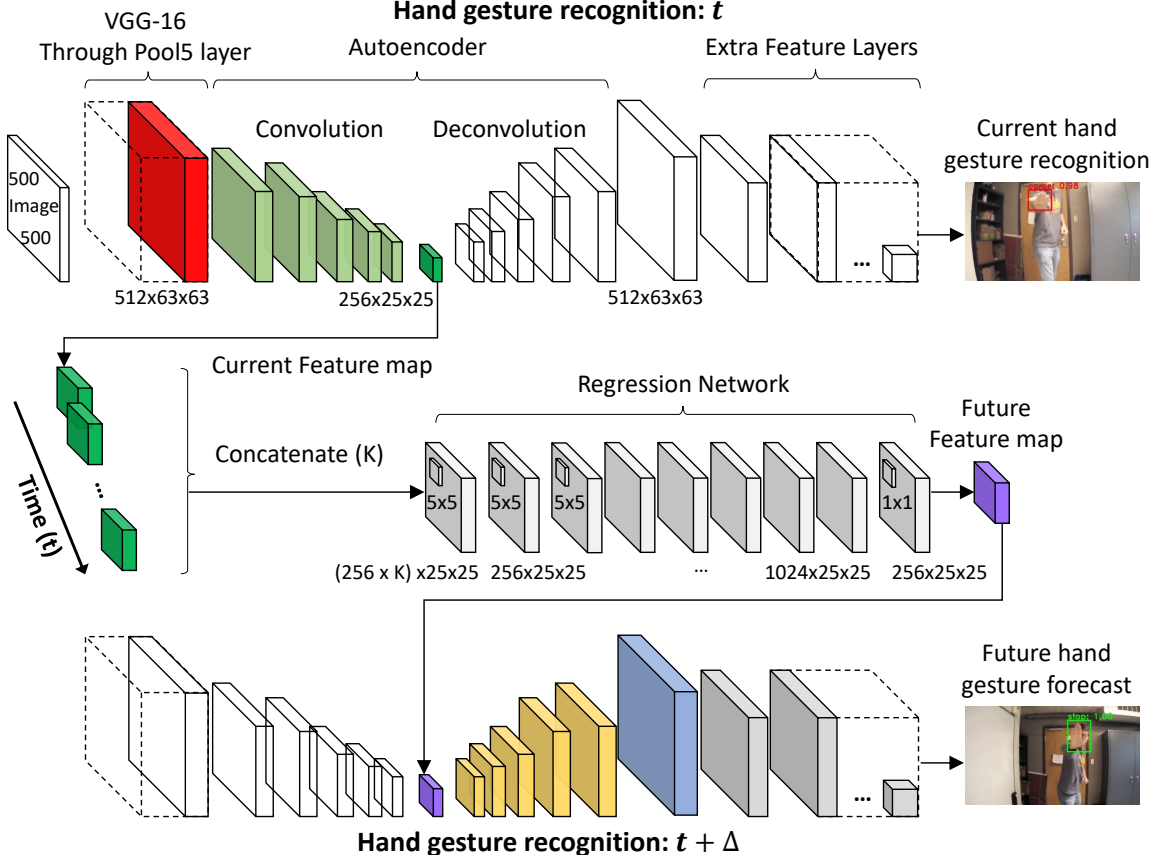


Figure 4.1: We employ the future regression network proposed in Section 3.1 to forecast a human partner’s future gestures.

#### 4.2.2 APPROACH: INTERACTIVE CONTROL BASED ON FORECAST-ING HAND GESTURES

We employ our fully convolutional future regression network which we described in Section 3.1 to forecast a human partner’s future gestures given the first few video frames of the gestures, and then perform a corresponding drone control command for each hand gesture. Figure 4.1 illustrates our adaption of that network for our this application scenario. As a reminder, the approach consists of three deep neural networks. The first row extracts visual features from the frames observed so far, the second row uses these to predict the visual features of a frame *one second in the future*, and then the third row uses these predicted features to classify the gesture in that “hallucinated” frame. The network in the second



Figure 4.2: Hand gestures for ‘come’ and ‘go’ to interact with a drone. Since we collect our interaction videos using a real drone and directly map drone control commands for each hand gesture, we can avoid the knowledge transfer problem in this scenario.

row of Figure 4.1 thus can be trained without gesture labels, although they are required to classify hand gestures.

### 4.2.3 EXPERIMENTAL RESULTS

#### Datasets

**Human-Drone Interaction (HDI) Videos:** We train the network on a new dataset of human-drone interaction videos consisting of 5 participants in two human-drone interaction scenarios (drone delivery and taking a self-portrait). We used the front-facing camera of the Parrot AR.Drone 2.0 for collecting the videos with a resolution of  $1280 \times 720$  at 30fps. Both scenarios were one-on-one interactions in which participants were told to direct the drone using five pre-defined hand gestures (selfie, stop, come, go, and change altitude). Each participant had 2-3 opportunities to interact with the drone for each scenario, and each interaction lasted about 1-3 minutes, yielding a total of 22 videos with 57,097 frames. We manually annotated 3,020 frames with ground-truth gesture labels, creating around 600 frames per gesture for training the network. Figure 4.2 shows the examples of ‘come’ and ‘go’ gestures in our dataset. Since we use the same drone for both collecting training

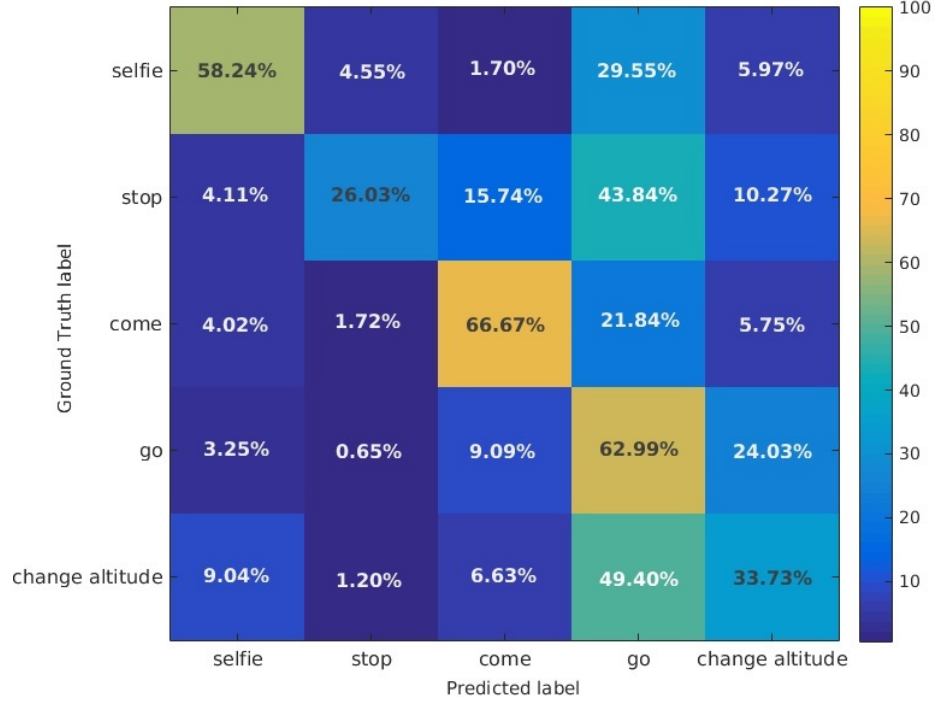


Figure 4.3: Confusion matrix of our gesture forecasting.

examples and testing, we can avoid the knowledge transfer problem in this scenario for generating drone control commands to interact with people.

### Hand Gesture Forecasting

We first trained the gesture recognition system on our Human-Drone Interaction videos to recognize hands in current frames, using only the annotated frames, and randomly splitting them into training (2,014 frames) and test (1,006). This achieved 99.1% accuracy in recognizing the current gesture on the test set.

Once the network in the first row of Figure 4.1 had been trained, we used it to extract scene representations from all frames of the videos. We then trained the future regression network in the second row of Figure 4.1 using the extracted scene features. Since this training process does not require any ground truth labels, we used all frames of the videos (two-thirds for training and the rest for evaluation). In the test phase, the regression network



was coupled with the hand gesture recognition network (third row) to predict *future* hand gestures given observations.

Figure 4.3 shows the confusion matrix of our hand gesture forecasting. We observe that our approach gives about 53.86% accuracy in predicting future human hand gestures one second ahead of time (compared to a random baseline of 20.0%), although it performed poorly on forecasting the ‘stop’ gesture. This result may be explained by the fact that participants used the ‘stop’ gesture in many different situations (i.e., when the drone was moving too far or approaching too close). The system also often confused ‘change altitude’ with the ‘go’ gesture, since pre-motions of these two gestures look similar.

## 4.3 HUMAN-ROBOT COLLABORATION WITH A HUMANOID ROBOT

### 4.3.1 INTRODUCTION

Our second application scenario is building the control function to move robot arms to imitate the collaborative behaviors learned from the human demonstration videos, using the proposed perception component which we described in Section 3.1.

Given a current video frame, we can predict approximate hand locations of the camera wearer based on the proposed perception component. On the basis of this prediction, our objective is to generate robot motor control commands to move the robot hands into the predicted camera wearer’s estimated future hand locations. However, a major obstacle with this approach is that the information provided by the perception component is insufficient for moving robot hands. A robot needs hand locations in the 3D world frame to move to, but the perception component only gives future hand location in 2D image coordinates.

We therefore construct a new regression network for mapping predicted 2-D human hand locations in image coordinates to the actual motor control commands for the robot arms. We use a humanoid robot (Baxter Research robot) as our hardware platform, then exper-

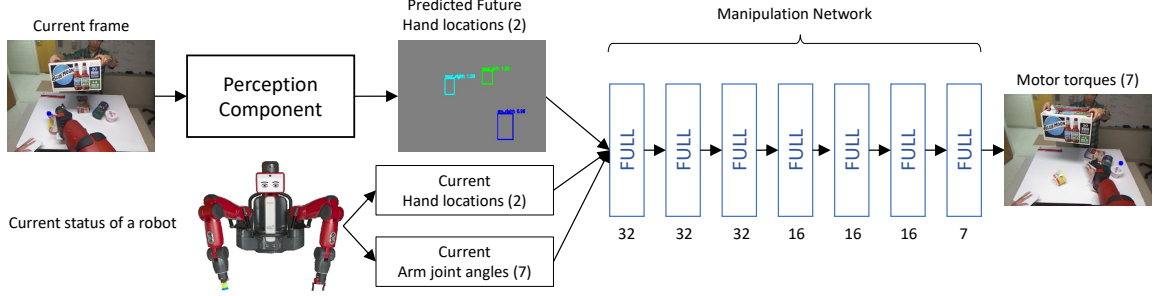


Figure 4.4: Robot manipulation component of our approach. It generates robot control commands given current robot joint state, current robot hand locations, and predicted future robot hand locations.

imentally confirm our approach through a user study with two human-robot collaboration scenarios.

### 4.3.2 APPROACH: MANIPULATION NETWORK FOR IMITATING HUMAN BEHAVIORS

We introduce a new regression network ( $m$ ) that maps the predicted 2-D camera wearer’s future hand locations in image coordinates to the actual motor control commands for moving the robot arms to that location.

The main assumption here is that a video frame from a robot’s camera will have a similar viewpoint to the first-person view point in our human videos, allowing us to take advantage of the trained model for predicting the “robot” hand locations in the future.

$$\hat{\mathbf{Y}}_{\mathbf{R}t} \simeq \hat{\mathbf{Y}}_t \quad (4.1)$$

where,  $\hat{\mathbf{Y}}_{\mathbf{R}t}$  represents robot hand locations and  $\hat{\mathbf{Y}}_t$  represents human hand locations at time  $t$ .

Figure 4.4 shows the robot manipulation network of our approach for this application scenario. With this assumption, our manipulation network ( $m$ ) predicts future robot joint states ( $\hat{\mathbf{Z}}_{t+\Delta}$ ) given current robot joint states ( $\hat{\mathbf{Z}}_t$ ), robot hand locations ( $\hat{\mathbf{Y}}_{\mathbf{R}t}$ ), and future

hand locations ( $\hat{\mathbf{Y}}_{\mathbf{R}t+\Delta}$ ) telling where the robot’s hands should move to. This network can be formulated with the below function:

$$\hat{\mathbf{Z}}_{t+\Delta} = m_{\theta}(\hat{\mathbf{Z}}_t, \hat{\mathbf{Y}}_{\mathbf{R}t}, \hat{\mathbf{Y}}_{\mathbf{R}t+\Delta}). \quad (4.2)$$

It predicts future robot joint states given current robot joint states, robot hand locations, and future hand locations telling where the robot’s hands should move to. It consists of seven fully connected layers having the following number of hidden units for each layer: 32, 32, 32, 16, 16, 16, 7. The weights ( $\theta$ ) of this network can be obtained in the same way used for our perception networks:

$$\theta^* = \arg \min_{\theta} \sum_{j,t} \|m_{\theta}(\hat{\mathbf{Z}}_t^j, \hat{\mathbf{Y}}_{\mathbf{R}t}^j, \hat{\mathbf{Y}}_{\mathbf{R}t+\Delta}^j) - \hat{\mathbf{Z}}_{t+\Delta}^j\|_2^2 \quad (4.3)$$

where  $\hat{\mathbf{Z}}_t^j$  indicates robot joint states at time  $t$  from training episode  $j$ , and  $\hat{\mathbf{Y}}_{\mathbf{R}t}^j$  represents robot hand locations at time  $t$  from training episode  $j$ .

The combination of our perception component and this manipulation network provides a real-time robotics system that takes raw video frames as its input and generates motor control commands for activity execution. Our manipulation network can be replaced with standard Inverse Kinematics, but our neural network-based model can potentially generate more natural arm movements by considering the desired location of the robot’s end-effectors as well as joint configuration sequences (i.e., unlabeled robot logs described in the next section).

### 4.3.3 EXPERIMENTAL RESULTS

We built a real-time robotics system based on the combination of our perception component and manipulation network that takes video frames from the robot camera and generates

robot control commands for performing the learned collaborative behaviors. The proposed system operates in slow real-time ( $\sim 100$  ms per frame) using one Nvidia Pascal Titan X GPU, and thus we can conduct real-time human-robot collaboration experiments based on our system. In this section, we will present our real-time robot experiments for evaluating our entire system with human participants.

## Datasets

**Unlabeled Robot Activity Log Files:** We prepared this dataset to train our robot manipulation component. It contains 50 robot log files, each of which records the robot’s hand positions  $(u, v)$  in an image plane and the robot’s corresponding joint angles at time  $t$ . We recorded these log files by having a human operator move the robot arms (i.e., the human grabbed the robot arms and moved them). We obtained such robot joint configuration sequences while moving the robot to cover possible arm motion during general human-robot interaction tasks. Here, we assume that the robot is supposed to operate in a similar environment during testing. Note that this data was not recorded in an interaction scenario (i.e., just the robot itself was moving), and no annotation regarding the activity or motion was provided. We used a Baxter research robot for recording these files. The Baxter has seven degree-of-freedom arms, so each file contains 9 variables for each arm (seven for joint angles and two for the robot hand positions in an image plane). In order to estimate the robot’s hand position in the image plane, we projected the 3-D positions of Baxter’s grippers into the image plane (based on camera calibration) and recorded the projected  $(u, v)$  positions with 7 joint angles at 30 Hz.

## Baselines

In addition to our approach (i.e., our perception component + manipulation component), we designed and implemented the following three baselines and compared our approach

Method	Task 1	Task 2	Average
Base SSD + Base control	$1.25 \pm 0.43$	$2.21 \pm 1.41$	$1.72 \pm 0.92$
Base SSD + Our control	$1.50 \pm 0.96$	$2.33 \pm 1.60$	$1.92 \pm 1.28$
Our perception + Base control	$2.33 \pm 1.18$	$2.25 \pm 1.36$	$2.29 \pm 1.27$
Ours	<b><math>3.17 \pm 1.40</math></b>	<b><math>3.42 \pm 1.61</math></b>	<b><math>3.29 \pm 1.50</math></b>

Table 4.1: The success level of our human-robot collaboration

with them: **(i) Base SSD + Base control** uses the baseline SSD with future annotations as a perception component and the base manipulation network that directs maps current hand locations in the image plane to the current seven joint angles. We trained the base manipulation network on the same robot activity log files described in Section 4.1. **(ii) Base SSD + Our control** uses SSD with future annotations as the perception component and our manipulation component to generate motor commands. **(iii) Our perception + Base control** used our perception component to predict future hand locations and the base control network for manipulation.

## Evaluation

We recruited a total of 12 participants (5 undergraduate and 7 graduate students) from the campus of Indiana University-Bloomington, and asked them to interact with our Baxter according to our two collaboration scenarios (clearing the table for a partner and preparing a trivet for a cooking pan) with four different robotics control algorithms (three baselines and our proposed approach). Each participant thus had 8 opportunities ( $2 \text{ scenarios} \times 4 \text{ different robotics systems}$ ) to interact with the robot and each interaction took about 2 - 3 minutes. These experiments were conducted in a realistic but controlled environment, and we randomly ordered the robotics control algorithms. Each subject therefore was exposed to the system in a different order.

After these interactions, participants were asked to complete a questionnaire about the

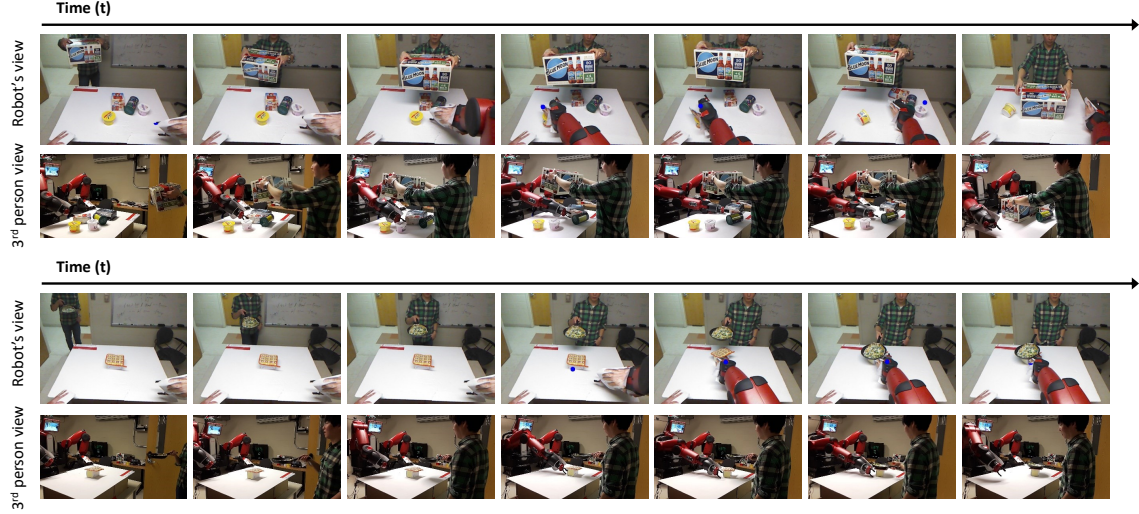


Figure 4.5: Qualitative results of our real-time robot experiments. There are two examples: clearing the table, and pushing the trivet toward the person. In each example, the first row shows the exact frames used as inputs to our robot (taken from a robot camera), and the second row shows the robot and the human from a 3rd person viewpoint. The frames were captured every second.

robot behaviors for each task. The questionnaire had two statements (one statement for each activity) with scales from 1 (totally do not agree) to 5 (totally agree) to express their impression of the robot behaviors: “I think the robot cleared the table to make a space for me” for task 1 and “I think the robot passed a trivet closer to me so that I can put the cooking pan on it” for task 2.

Table 4.1 shows the results, indicating that our participants evaluated our approach performed better on both tasks. We received a higher average score of 3.29 compared to all the baselines (1.72, 1.92, and 2.29) from the participants. Examples of our real-time robot experiments with human subjects are illustrated in Fig. 4.5.

## 4.4 A CASE STUDY FOR BUILDING A SOCIABLE ROBOT

### 4.4.1 INTRODUCTION

So far we have focused on methods to learn human activities from demonstration videos, and how to transfer the learned knowledge for building intelligent systems for robots. Let

us now turn to the topic of building “interactive” systems.

In this section, we present a case study of designing a sociable drone for natural human-drone interaction. We explore the critical factors for successful human-drone interaction in a social scenario through a focus group interview, and discuss our findings and suggestions to build a sociable drone.

Drones have developed rapidly in the past few years and have been applied to many social scenarios, such as delivery services [3], ping-ping play [81], and jogging companions [40]. However, despite the growing interest, only a few studies have tried to discover how people interact with them [105]. Furthermore, most of this work focuses on gesture recognition to control drones [18, 79], despite that there are many other important factors such as the drone’s movements, appearance, and distance that could influence whether and how people engage in the interaction [28].

In this case study, we aim to explore the designable factors that could make a drone more acceptable in social scenarios. Few studies have investigated design considerations for people’s interaction with a flying robot, so, many interesting research questions still remain, such as how drone’s behavior and its appearance change people’s perception of it. Therefore, we conduct a user study to observe how people respond to drones when it is flying at a close distance to interact with them.

#### **4.4.2 APPROACH: CARD SORTING WITH FOCUS GROUP INTERVIEW**

##### **Research Method Design**

To understand people’s impressions and feelings of a moving drone, we conducted a user study with 4 participants (3 females, 1 male) in a simulated domestic environment. Each participant had a chance to interact with a drone (Parrot AR Drone 2.0) according to



Figure 4.6: An example of our experimental setup with the drone. We place a small object on top of the drone for our experimental scenario and manually controlled the drone during the interaction. Each participant was finally asked to pick up the object from the drone.

our interaction scenario in which people send a drone to deliver small objects to a specific person or area for a social event. All participants were recruited via email advertisement and received no remuneration.

During the interaction, we manually controlled the drone. First, we attached a small plate to the top of the drone and put a small object as the object to be delivered to the participant. After that we lifted the drone about five feet away from the participant, and then we moved it to approach the subject until it was close enough for the subject to reach the object from the drone. Once the the drone approached the subject, we finally asked the person to pick up the object, and then we moved the drone back to the starting position. The whole scenario for each participant lasted about 5 minutes on average. To protect people and the drone, we placed a foam hull (with black and green camouflage patterns) on the drone and stopped the experiment if requesting.

After the interaction with the drone, we conducted a card sorting session, in which we asked participants to write words to describe the drone and their feelings about it on



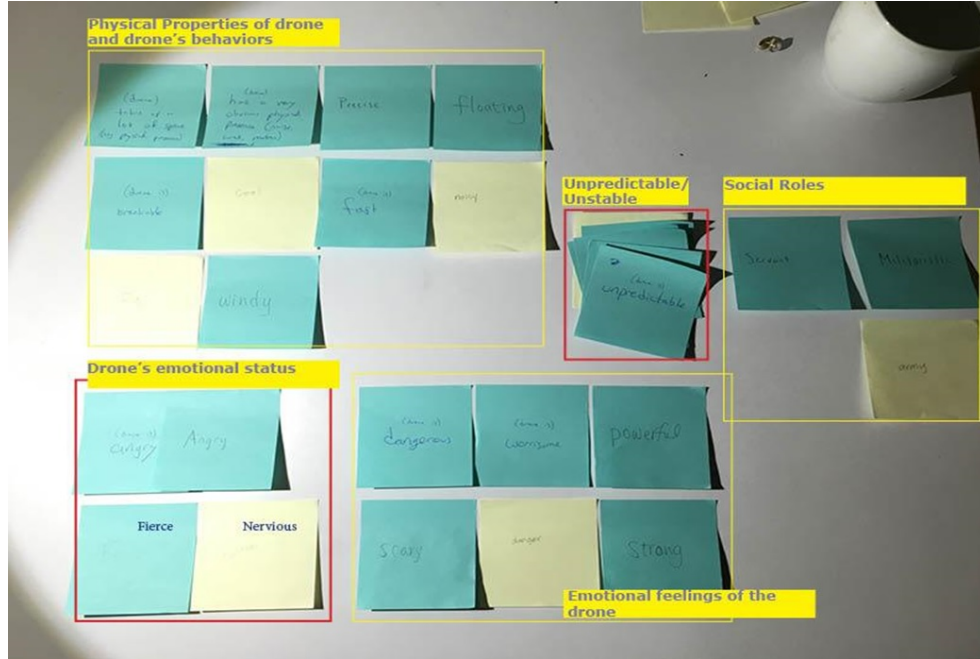


Figure 4.7: We conducted a card sorting session and categorized the collected words into four different groups.

notes. We then conducted a focus group interview with our participants to understand their experiences of interacting with the drone. Interview questions included topics such as any discomfort they felt during the interaction and why, how they felt about the interaction distance and the appearance of the drone, etc. Each interview lasted about 40 minutes.

### 4.4.3 FINDINGS

#### Card Sorting Session

We categorized the collected words into four groups after conducting the card sorting session: drone's physical properties and behaviors, social roles, drone's emotional status, and participant's feelings of the interaction.

Overall, participants expressed negative impressions about the drone in the card sorting session. We found a significant number of notes as indicating "unpredictable" or "unstable" as the most notable traits of the drone. They also described the drone as noisy, windy, break-

able, and fast. In addition, we found some of them described the drone’s flying behaviors as a reflection of the drone’s emotional status (e.g. angry, nervous and fierce). Participants also made interpretations of social roles regarding drone’s appearance and behavior. They thought the drone was for “military” and “army,” or looked like a “servant.”

### **Focus Group Interview**

We were able to better understand why they felt this negative emotions by following the focus group interview. Some participants thought that the drone was intrusive since it moved too fast and sometimes too close to them. Furthermore, some mentioned that they preferred to approach the drone themselves once the drone stopped at a certain distance for picking up the delivered object. These results are likely to be related to the unpredictability of the drone’s behavior, since the drone seemed most stable when it was hovering at a static position.

The participants also mentioned the appearance of the drone, and that it reminded them of military. This may be explained by the fact that we used the foam hull which had black and green camouflage patterns for the experiments. They expected the drone to adopt modern design, and have a “cool” appearance. One participant even suggested saying the drone business look to make it appear more reliable. Noticeably, when discussing preferable appearances, participants rejected the suggestion of “cuteness.” They felt it was too counter-intuitive since cuteness did not match with drone’s behaviors.

### **4.4.4 DISCUSSION**

Our findings reflected participants’ concerns about potential dangers when interacting with a drone they could not predict. Participants also stated that they felt the drone intruded on them if it approached too closely without their consent. Results also indicated that people felt uncomfortable and even frightened by a drone’s appearance. To mitigate this issue, we

present the several design suggestions to encourage better human-drone interaction.

### **1. Showing Your Intentions**

A drone needs to show its intention to make the interaction comfortable and secure. Such social cues could be particularly important in the interaction with a drone, since its behavior could be more unpredictable than mobile robots because it moves in a three dimensional space. Therefore, having a drone shows its intention to subjects such as the intended direction of motion and/or speed would be desirable.

### **2. Respecting to Personal Space**

Our interview findings imply that people want to have a sense of control of the interaction process, especially in their personal space. Therefore, designers should leave enough space and give enough controls to users when designing the human-drone interaction procedure.

### **3. A Friendly Appearance**

In our interview, participants mentioned that the original appearance reminded them of military. Therefore, designers should try to generate a friendlier appearance. For example, participants mentioned they would like a modern design or business look. Designers should follow the design rules of transforming the drones into the objects we normally encounter in daily life.

## CHAPTER 5

### CONCLUSION AND FUTURE WORK

#### 5.1 THESIS SUMMARY AND DISCUSSION

In this thesis we have addressed the problem of activity learning from human demonstrations for building intelligent and interactive systems. Our approach is based on the research paradigm of robot learning from demonstration (LfD) with video demonstrations, which attempts to enable an intelligent agent to autonomously learn a new skill by watching demonstrations itself. We used Convolutional Neural Network (CNN) based approaches to capture the spatio-temporal structure of the demonstration videos and interpreted the demonstrated human actions based on hands movements. By using hands as medium to model human activities, we were able to understand the demonstrated human behaviors in the videos and reduce the correspondence problem for building diverse intelligent systems, from interactive robots to intelligent piano tutoring systems.

We also have proposed to use wearable cameras to collect human activity videos in order to take advantage of the information in first-person perspectives, such as the camera wearer’s intentions and experiences, and the in viewpoint similarity with a robot’s visual input. We showed that these types of videos helped to reduce cost of transferring knowledge from an intelligent agent’s perception system to its control system.

In Chapter 2, we reviewed background and related work on LfD while focusing on learning collaborative tasks. We surveyed several attempts to build better communication

frameworks between robots and humans using various signals from non-verbal cues like facial expressions to human/robot gestures and motions. We also explored another interesting research method, Interactive/Active learning, which considers a robot as an active partner that can provide feedback to the human user. We then wrapped up this chapter by briefly reviewing the common theoretical issues and the challenges.

In Chapter 3, we presented our CNN based perception components for observing various demonstration videos. We first introduced a novel fully convolutional network to analyze human hand movements in the videos of two people collaborating. The core idea of the proposed network was to extract scene representations from the intermediate layers of the network and use them to capture spatial-temporal information of the demonstrated human behaviors. We experimentally showed that our approach was able to capture the temporal structure of human activity and reliably predict the future locations of hands.

In the next section, we extended the proposed network by adding a separate temporal stream with motion-domain features to the original network for analyzing state changes of objects in videos of daily activity. The aim of this study was to understand how the objects near us will be moved according to daily activities. In the experimental results, we confirmed that our object forecast approach reliably estimated the future locations of hands and objects in the video and showed that it significantly outperforms the state-of-the-art future object presence forecast method on a public dataset.

In the last section of this chapter, we proposed another novel audio-visual fusion network for analyzing people playing a piano. The proposed network was designed to determine which notes on a piano are being played at any moment in time, and identify fingers used to press notes. We formulated this problem as object detection with multi-task learning, and demonstrated that our approach was able to detect pressed piano keys and the piano player’s fingers with high accuracy.

In Chapter 4, we presented two different methods for providing proper responses to human activities beyond perceiving them. We first showed a relatively simple method which directly maps the robot control commands to each human hand gesture for controlling a drone, and then introduced a new regression network to generate robot control commands for moving robot arms on a humanoid robot. Our study results confirmed that the robot was able to infer its motor control plan for generating learned collaborative behaviors based on the combination of perception and manipulation networks.

Finally, we reported our case study result to discuss building “interactive” systems. The purpose of this case study was to explore the critical factor for successful human-drone interaction in social scenarios. We conducted a user study to understand people’s impressions and feelings of a moving drone, and then reported the participants’ concerns about interacting with it. We then suggested some design implications to migrate their concerns and make drone more acceptable in social scenarios.

## **5.2 FUTURE WORK**

Although this dissertation addressed many important issues of learning human activities for building intelligent and interactive systems, the study is limited by not handling many human-centric and pedagogical issues for our intelligent piano tutoring systems. It also has many technical limitations. For example, a robot cannot semantically understand human behaviors based on our approaches even though it is able to imitate the human actions. Moreover, our approaches also have limited ability to handle the knowledge transfer problem, since they still require a separate training phase with a hardware platform to employ the learned knowledge from the video demonstrations. However, since the scope of this thesis is broad, we focused on building better perception components in this dissertation.

As future work, we first plan to revise some of our perception components to employ

some additional information to improve the recognition performance. For example, we have not utilized the temporal information of activities to obtain the pianist’s finger movements for our intelligent piano tutoring system in Chapter 3.

To mitigate the knowledge transfer problem in robot activity learning, we plan to use both human videos and robot videos which not only contain human hands, but also the robot hands, for more accurate scene representations.

Finally, we plan to conduct a user study to determine the effects of early recognition for human-robot interaction. Since most papers so far present just the technical approaches without investigating the effects of early recognition on human-robot interaction using actual subjects, we believe our study will explore people’s attitudes towards early recognition and the limitations of the current state-of-the-art future forecasting approaches.

## BIBLIOGRAPHY

- [1] Mohammad Akbari and Howard Cheng. Real-time piano music transcription based on computer vision. *IEEE Transactions on Multimedia*, 2015.
- [2] Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J Davison. Kaze features. In *European Conference on Computer Vision (ECCV)*, 2012.
- [3] Amazon.com, Inc. Amazon prime air. [online] <http://www.amazon.com/primeair>, 2013.
- [4] Heni Ben Amor, David Vogt, Marco Ewerton, Erik Berger, Bernhard Jung, and Jan Peters. Learning responsive robot behavior by imitation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013.
- [5] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 2009.
- [6] Sven Bambach, Stefan Lee, David J. Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.
- [7] Mathieu Barthet, Amélie Anglade, Gyorgy Fazekas, Sefki Kolozali, and Robert Macrae. Music recommendation for music learning: Hotttabs, a multimedia guitar tutor. In *The ACM Conference Series on Recommender Systems (RecSys) Workshop on Music Recommendation and Discovery*, 2011.



- [8] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 2013.
- [9] Aude Billard, Sylvain Calinon, Ruediger Dillmann, and Stefan Schaal. Robot programming by demonstration. In *Springer Handbook of Robotics*. Springer, 2008.
- [10] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2001.
- [11] Paulo Vinicius Koerich Borges, Nicola Conci, and Andrea Cavallaro. Video-based human behavior understanding: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2013.
- [12] Cynthia Breazeal, Andrew Brooks, Jesse Gray, Guy Hoffman, Cory Kidd, Hans Lee, Jeff Lieberman, Andrea Lockerd, and David Mulanda. Humanoid robots as cooperative partners for people. *International Journal of Humanoid Robots*, 2004.
- [13] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2005.
- [14] Maya Cakmak, Crystal Chao, and Andrea L Thomaz. Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development*, 2010.
- [15] Maya Cakmak and Andrea L Thomaz. Designing robot learners that ask good questions. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2012.

- [16] Sylvain Calinon and Aude Billard. Teaching a humanoid robot to recognize and reproduce social cues. In *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2006.
- [17] Sylvain Calinon, Paul Evrard, Elena Gribovskaya, Aude Billard, and Abderrahmane Kheddar. Learning collaborative manipulation tasks by demonstration using a haptic interface. *International Conference on Advanced Robotics*, 2009.
- [18] Jessica R Cauchard, Kevin Y Zhai, James A Landay, et al. Drone & me: an exploration into natural human-drone interaction. In *ACM International joint conference on pervasive and ubiquitous computing*, 2015.
- [19] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.
- [20] Sonia Chernova and Andrea L Thomaz. Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2014.
- [21] Jonathan Chow, Haoyang Feng, Robert Amor, and Burkhard C Wünsche. Music education using augmented reality with a head mounted display. In *Australasian User Interface Conference*, 2013.
- [22] Roger B Dannenberg, Marta Sanchez, Annabelle Joseph, Peter Capell, Robert Joseph, and Ronald Saul. A computer-based multi-media tutor for beginning piano students. *Journal of New Music Research*, 1990.
- [23] Arnaud Dessein, Arshia Cont, and Guillaume Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

- [24] Christian Dittmar, Estefanía Cano, Jakob Abeßer, and Sascha Grollmisch. Music Information Retrieval Meets Music Education. In *Multimodal Music Processing*, 2012.
- [25] Christian Dittmar and Daniel Gärtner. Real-time transcription and separation of drum recordings based on nmf decomposition. In *DAFx*, 2014.
- [26] Anca Dragan and Siddhartha Srinivasa. Familiarization to robot motion. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2014.
- [27] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2013.
- [28] Brittany A Duncan and Robin R Murphy. Comfortable approach distance with small unmanned aerial vehicles. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2013.
- [29] Marco Ewerton, Gerhard Neumann, Rudolf Lioutikov, Heni Ben Amor, Jan Peters, and Guilherme Maeda. Learning multiple collaborative tasks with a mixture of interaction primitives. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.
- [30] Chenyou Fan, Jangwon Lee, and Michael S Ryoo. Forecasting hand and object locations in future frames. *European Conference on Computer Vision (ECCV) Workshop on Anticipating Human Behavior*, 2018.
- [31] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

- [32] Alireza Fathi, Ali Farhadi, and James M. Rehg. Understanding egocentric activities. In *International Conference on Computer Vision (ICCV)*, 2011.
- [33] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [34] Nadia Figueroa, Ana Lucia Pais Ureche, and Aude Billard. Learning complex sequential tasks from demonstration: A pizza dough rolling case study. In *ACM/IEEE International Conference on Human Robot Interaction (HRI)*, 2016.
- [35] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.
- [36] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.
- [37] James Jerome Gibson. The senses considered as perceptual systems. *Houghton Mifflin*, 1966.
- [38] Michael A Goodrich and Alan C Schultz. Human-robot interaction: a survey. *Foundations and trends in human-computer interaction*, 2007.
- [39] Ilaria Gori, JK Aggarwal, Larry Matthies, and Michael S Ryoo. Multi-type activity recognition in robot-centric scenarios. *IEEE Robotics and Automation Letters (RA-L)*, 2016.
- [40] Eberhard Graether and Florian Mueller. Joggobot: a flying robot as jogging companion. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2012.

- [41] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [42] Charles Louis Hanon. *The virtuoso pianist: in sixty exercises for the piano*, volume 1. G. Schirmer, 1911.
- [43] Karol Hausman, Scott Niekum, Sarah Osentoski, and Gaurav S Sukhatme. Active articulation model estimation through interactive perception. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.
- [44] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck. Onsets and frames: Dual-objective piano transcription. *arXiv preprint arXiv:1710.11153*, 2017.
- [45] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Overview of mini-batch gradient descent. *Neural Networks for Machine Learning - Lecture 6a, University of Toronto*, 2012.
- [46] Simon Holland. Artificial intelligence in music education: A critical review. In *Readings in Music and Artificial Intelligence*, 2000.
- [47] Paul VC Hough. Method and means for recognizing complex patterns, December 18 1962. US Patent 3,069,654.
- [48] Yuri A. Ivanov and Aaron F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2000.

- [49] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-RNN: Deep learning on spatio-temporal graphs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [50] David Johnson, Isabelle Dufour, Daniela Damian, and George Tzanetakis. Detecting pianist hand posture mistakes for virtual piano tutoring. In *International Computer Music Conference (ICMC)*, 2016.
- [51] Geetanjali Vinayak Kale and Varsha Hemant Patil. A study of vision based human motion recognition and analysis. *International Journal of Ambient Computing and Intelligence (IJACI)*, 2016.
- [52] Kris M. Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
- [53] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European Conference on Computer Vision*. Springer, 2012.
- [54] W Bradley Knox, Peter Stone, and Cynthia Breazeal. Training a robot via human feedback: A case study. In *International Conference on Social Robotics*. Springer, 2013.
- [55] Jonas Koenemann, Felix Burget, and Maren Bennewitz. Real-time imitation of human whole-body motions by humanoids. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.
- [56] Hema Koppula and Ashutosh Saxena. Physically-grounded spatio-temporal object affordances. In *European Conference on Computer Vision (ECCV)*, 2014.

- [57] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 2016.
- [58] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 2013.
- [59] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 2005.
- [60] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [61] Przemyslaw Lasota, Stefanos Nikolaidis, and Julie A Shah. Developing an adaptive robotic assistant for close proximity human-robot collaboration in space. In *AIAA Infotech@ Aerospace Conference*, 2013.
- [62] Jangwon Lee and Michael S Ryoo. Learning robot activities from first-person human videos using convolutional future regression. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.
- [63] Jangwon Lee, Haodan Tan, David Crandall, and Selma Šabanović. Forecasting hand gestures for human-drone interaction. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI), Late-Breaking Report*, 2018.
- [64] Jangwon Lee, Jingya Wang, David Crandall, Selma Šabanović, and Geoffrey Fox. Real-time, cloud-based object detection for unmanned aerial vehicles. In *IEEE International Conference on Robotic Computing (IRC)*, 2017.

- [65] Kyuhwa Lee, Yanyu Su, Tae-Kyun Kim, and Yiannis Demiris. A syntactic approach to robot imitation learning using probabilistic activity grammars. *Robotics and Autonomous Systems*, 2013.
- [66] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 2016.
- [67] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 2016.
- [68] Dawen Liang, Minshu Zhan, and Daniel PW Ellis. Content-aware collaborative music recommendation using pre-trained neural networks. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2015.
- [69] Hui Liang, Jin Wang, Qian Sun, Yong-Jin Liu, Junsong Yuan, Jun Luo, and Ying He. Barehanded music: real-time hand interaction for virtual piano. In *ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 2016.
- [70] Hongyi Liu and Lihui Wang. Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics*, 2017.
- [71] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, ChengYang Fu, and Alexander Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, 2016.
- [72] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *International Conference on Learning Representations (ICLR)*, 2017.



- [73] David G Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision (ICCV)*, 1999.
- [74] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [75] Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [76] Wei-Chiu Ma, De-An Huang, Namhoon Lee, and Kris M. Kitani. Forecasting interactive dynamics of pedestrians with fictitious play. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [77] Jim Mainprice and Dmitry Berenson. Human-robot collaborative manipulation planning using early prediction of human motion. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013.
- [78] Dipendra K Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. *The International Journal of Robotics Research*, 2016.
- [79] Kensho Miyoshi, Ryo Konomura, and Koichi Hori. Above your hand: direct and natural interaction with aerial robot. In *ACM SIGGRAPH Emerging Technologies*, 2014.
- [80] Anahita Mohseni-Kabir, Charles Rich, Sonia Chernova, Candace L Sidner, and Daniel Miller. Interactive hierarchical task learning from a single demonstration. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2015.

- [81] Mark Müller, Sergei Lupashin, and Raffaello D’Andrea. Quadrocopter ball juggling. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2011.
- [82] Katharina Mülling, Jens Kober, Oliver Kroemer, and Jan Peters. Learning to select and generalize striking movements in robot table tennis. *International Journal of Robotics Research*, 2013.
- [83] Bilge Mutlu. Designing embodied cues for dialog with robots. *AI Magazine*, 2011.
- [84] Chrystopher L Nehaniv, Kerstin Dautenhahn, et al. The correspondence problem. *Imitation in animals and artifacts*, 2002.
- [85] Truong-Huy Dinh Nguyen, David Hsu, Wee-Sun Lee, Tze-Yun Leong, Leslie Pack Kaelbling, Tomas Lozano-Perez, and Andrew Haydn Grant. Capir: Collaborative action planning with intention recognition. *arXiv preprint arXiv:1206.5928*, 2012.
- [86] Scott Niekum, Sarah Osentoski, George Konidaris, and Andrew G Barto. Learning and generalization of complex tasks from unstructured demonstrations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.
- [87] Stefanos Nikolaidis and Julie Shah. Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2013.
- [88] Sergio Oramas, Oriol Nieto, Francesco Barbieri, and Xavier Serra. Multi-label music genre classification from audio, text, and images using deep features. *arXiv preprint arXiv:1707.04916*, 2017.

- [89] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [90] Graham Percival, Ye Wang, and George Tzanetakis. Effective use of multimedia for computer-assisted musical instrument tutoring. In *ACM International workshop on Educational multimedia and multimedia education*, 2007.
- [91] Alfonso Perez-Carrillo, Josep-Lluís Arcos, and Marcelo Wanderley. Estimation of guitar fingering and plucking controls based on multimodal analysis of motion, audio and musical score. In *International Symposium on Computer Music Multidisciplinary Research*, 2015.
- [92] Leah Perlmutter, Eric Kernfeld, and Maya Cakmak. Situated language understanding with human-like and visualization-based transparency. In *Robotics: Science and Systems (RSS)*, 2016.
- [93] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.
- [94] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [95] Michalis Raptis, Iasonas Kokkinos, and Stefano Soatto. Discovering discriminative action parts from mid-level video representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [96] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

- [97] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.
- [98] M. S. Ryoo, Thomas J. Fuchs, Lu Xia, J. K. Aggarwal, and Larry Matthies. Robot-centric activity prediction from first-person videos: What will they do to me? In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2015.
- [99] Michael S Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.
- [100] Allison Sauppé and Bilge Mutlu. The social impact of a robot co-worker in industrial settings. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.
- [101] Tianmin Shu, Michael S. Ryoo, and Song-Chun Zhu. Learning social affordance for human-robot interaction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.
- [102] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon. An end-to-end neural network for polyphonic piano music transcription. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 2016.
- [103] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.
- [104] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

- [105] Daniel J Szafr. Human interaction with assistive free-flyers. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI)s*, 2014.
- [106] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [107] Leila Takayama, Doug Dooley, and Wendy Ju. Expressing thought: improving robot readability with animation principles. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2011.
- [108] Yoshinari Takegawa, Tsutomu Terada, and Shojiro Nishio. Design and implementation of a real-time fingering detection system for piano performance. In *International Computer Music Conference (ICMC)*, 2006.
- [109] Haodan Tan, Jangwon Lee, and Gege Gao. Human-drone interaction: Drone delivery & services for social events. In *Conference on Designing Interactive Systems (DIS), Work-in-Progress*, 2018.
- [110] Stefanie Tellex, Ross A Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. Asking for help using inverse semantics. In *Robotics: Science and systems (RSS)*, 2014.
- [111] Stefanie A Tellex, Thomas Fleming Kollar, Steven R Dickerson, Matthew R Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI Conference on Artificial Intelligence*, 2011.
- [112] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE international conference on computer vision (ICCV)*, 2015.

- [113] Richard Vogl, Matthias Dorfer, Gerhard Widmer, and Peter Knees. Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks. In *the International Society for Music Information Retrieval Conference (ISMIR)*, 2018.
- [114] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations with unlabeled video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [115] Jacob Walker, Abhinav Gupta, and Martial Hebert. Patch to the future: Unsupervised visual prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [116] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [117] Di Wu and Ling Shao. Silhouette analysis-based action recognition via exploiting human poses. *IEEE Transactions on Circuits and Systems for Video Technology*, 2013.
- [118] Caiming Xiong, Nishant Shukla, Wenlong Xiong, and Song-Chun Zhu. Robot learning with a spatial, temporal, and causal and-or graph. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.
- [119] Junji Yamato, Jun Ohya, and Kenichiro Ishii. Recognizing human action in time-sequential images using hidden markov model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1992.

- [120] Yezhou Yang, Yi Li, Cornelia Fermüller, and Yiannis Aloimonos. Robot learning manipulation action plans by “watching” unconstrained videos from the world wide web. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2015.
- [121] Daqing Yi and Michael A Goodrich. Supporting task-oriented collaboration in human-robot teams using semantic-based path planning. In *Unmanned Systems Technology XVI*, 2014.
- [122] Jun Yin, Ye Wang, and David Hsu. Digital violin tutor: an integrated system for beginning violin learners. In *ACM international conference on Multimedia*, 2005.
- [123] Yasuyoshi Yokokohji, Yuki Kitaoka, and Tsuneo Yoshikawa. Motion capture from demonstrator’s viewpoint and its application to robot teaching. *Journal of Field Robotics*, 2005.
- [124] Bingjun Zhang and Ye Wang. Automatic music transcription using audio-visual fusion for violin practice in home environment. In *Technical Report TRA7/09, Shool of Computing, National University of Singapore*, 2009.
- [125] Fangyi Zhang, Jürgen Leitner, Michael Milford, Ben Upcroft, and Peter Corke. Towards vision-based deep reinforcement learning for robotic motion control. *arXiv preprint arXiv:1511.03791*, 2015.
- [126] Xinquan Zhou and Alexander Lerch. Chord detection using deep learning. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2015.
- [127] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. *arXiv preprint arXiv:1707.07012*, 2017.

## CURRICULUM VITAE

### Jangwon Lee

School of Informatics, Computing, and Engineering  
Indiana University Bloomington  
Email: [leejang@indiana.edu](mailto:leejang@indiana.edu)  
Homepage: <http://homes.soic.indiana.edu/leejang/>

### EDUCATION

- Ph.D. Informatics, Indiana University, Bloomington, 2018.
- M.S. Electrical and Computer Engineering, Sungkyunkwan University, 2008.
- B.S. Electronic and Electrical Engineering, Sungkyunkwan University, 2006.

### EMPLOYMENT

- Indiana University, Research Assistant/Associate Instructor, 2013–2018.
- NASA Jet Propulsion Laboratory, Summer Research Intern, 2016–2016.
- Samsung Electronics, Software Engineer, 2010–2013.
- Samsung Digital Imaging, Research Engineer, 2009–2010.
- Samsung Techwin, Assistant Research Engineer, 2008–2009.
- Sungkyunkwan University, Research Assistant, 2006–2008.

### PUBLICATIONS

#### PEER-REVIEWED CONFERENCE PAPERS

1. Jangwon Lee and Michael S.Ryoo. Learning Robot Activities from First-Person Human Videos Using Convolutional Future Regression. In *IEEE/RSJ International Con-*



- ference on Intelligent Robots and Systems (IROS)*, Sep 2017.
2. Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David Crandall, and Michael S.Ryoo. Identifying First-person Camera Wearers in Third-person Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.
  3. Jangwon Lee, Jingya Wang, David Crandall, Selma Šabanović, and Geoffrey Fox. Real-Time, Cloud-Based Object Detection for Unmanned Aerial Vehicles. In *IEEE International Conference on Robotic Computing (IRC)*, Apr 2017.
  4. Hyunjun Kim, Jangwon Lee, and Sukhan Lee. Environment adaptive 3d object recognition and pose estimation by cognitive perception engine. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, Dec 2009.
  5. Jangwon Lee, Dongwook Shin, Hunsue Lee, and Sukhan Lee. Study on behavioral personality of a service robot to make more convenient to customer. In *The 16th IEEE International Symposium on Robot and Human interactive Communication (RO-MAN)*, Aug 2007.
  6. Seung-Min Baek, Jangwon Lee, Hunsue Lee, Dongwook Shin, and Sukhan Lee. Information integration and mission selection to accomplish dependable perception for service robot. In *The 13th International Conference on Advanced Robotics (ICAR)*, Aug 2007.
  7. Hunsue Lee, Jangwon Lee, Jaewoong Kim, and Sukhan Lee. Security service robot in ubiquitous environment based on cognitive robotic engine. In *The 1st International Conference of Ubiquitous Information Technology and Applications (ICUT)*, Feb 2007.
  8. Dongwook Shin, Jangwon Lee, Hun-Sue Lee, Sukhan Lee, Young-Jo Cho, and Su-Young Chi. Robot personality from perceptual behavior engine: An experimental

study. In *The 3rd International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, Oct 2006.

## **EXTENDED ABSTRACTS IN CONFERENCES AND WORKSHOPS**

1. Chenyou Fan, Jangwon Lee, and Michael S. Ryoo. Forecasting Hand and Object Locations in Future Frames. In *European Conference on Computer Vision (ECCV) Workshop on Anticipating Human Behavior*, Sep 2018
2. Haodan Tan, Jangwon Lee and Gege Gao. Human-Drone Interaction: Drone Delivery & Services for Social Event. In *DESIGNING INTERACTIVE SYSTEMS (DIS), Provocations and Works-in-Progress (PWiP)*, Jun 2018
3. Jangwon Lee, Haodan Tan, Selma Šabanović, and David Crandall. Forecasting Hand Gestures for Human-Drone Interaction. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI), Late-Breaking Reports*, Mar 2018.
4. Jangwon Lee and Michael S. Ryoo. Learning Robot Activities from First-person human Videos Using Convolutional Future Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Deep Learning for Robotic Vision (DLRV)*, Jul 2017
5. Jangwon Lee and Michael S. Ryoo. Learning Robot Activities from First-Person Human Videos Using Convolutional Future Regression. In *IEEE International Conference on Robotics and Automation (ICRA), Late Breaking Results Poster Session*, May 2017.

## **BOOK CHAPTERS**

1. Sukhan Lee, Seung-Min Baek, and Jangwon Lee. Cognitive robotic engine: Behavioral perception architecture for human-robot interaction. In *Human Robot Interaction*, chapter 13. Nilanjan Sarkar (Ed.), ISBN: 978-3-902613-13-4, InTech, Sep 2007

## PATENTS

1. Sungwook Lee and Jangwon Lee. Method and apparatus for photographing an image in a user device, 2013. US Patent 9,596,412.
2. Jangwon Lee. Digital photographing apparatus, method of controlling the same, and recording medium having recorded thereon program for executing the method, 2012. US Patent 8,872,959.
3. Eunyoung Kim and Jangwon Lee. Method and apparatus for capturing moving picture, 2012. US Patent App. 13/282,761.
4. Eunyoung Kim and Jangwon Lee. Apparatus for processing digital image and thereof method, 2011. Korea Patent Publication Number: 10-2011-0087595.
5. Jangwon Lee. Apparatus and method for image processing using security function, 2011. US Patent 8,482,633
6. Jangwon Lee. Digital image signal processing method, medium for recording the method, and digital image signal processing apparatus, 2010. US Patent 9,426,359.
7. Jangwon Lee. Photographing control method and apparatus using stroboscope, 2010. Korea Patent Publication Number: 10-2010-0077715.
8. Jangwon Lee. Digital camera supporting intelligent self-timer mode and method of controlling the same, 2010. US Patent 8,711,232.
9. Sukhan Lee, Seung-Min Baek, Jaihun Lee, and Jangwon Lee. System and method for real-time object recognition and pose estimation using in-situ monitoring, 2009. US Patent 8,503,760.

## AWARDS AND SCHOLARSHIPS

- Best Paper Award, CVPR Workshop on Deep Learning for Robotic Vision, 2017
- Travel Grant, CVPR Workshop Deep Learning for Robotic Vision, 2017
- Fellowship, four years of tuition and stipend, Indiana University, 2013

- Brain Korea 21 Scholarship, Sungkyunkwan University, 2008